JOHDATUS KEMINFORMATIIKKAAN

Intensiivinen itseopiskeluopas aloittelijoille

David Wild

Edumendo
ADVANCED LEARNING AND VISUALIZATION

JOHDATUS KEMINFORMATIIKKAAN

INTENSIIVINEN ITSEOPISKELUOPAS ALOITTELIJOILLE

David Wild

Suomentanut
Johannes Pernaa

KUSTANTAJA
Edumendo Oy | www.edumendo.fi

KUVAOIKEUDET: Adobe Stock: 49; Freepik: 89, 105, 109; Harryarts / Freepik: 77; macrovector_official / Freepik: 5 ja 122; starline / Freepik: I, II, 3, 9, 54, 63, 71, 98; Katemangostar / Freepik: 54; vector_corp / Freepik: 16, 30, 37; rawpixel.com / Freepik: kansikuva

PAINETUN VERSION VALMISTAJA
Books on Demand GmbH, Norderstedt, Saksa

TUNNISTEET
ISBN 978-952-94-4390-1 (nid.)
ISBN 978-952-94-4391-8 (PDF)
DOI 10.31885/9789529443918

VERSIO: 1.0 | 2.2.2021 |

SUOMENNOSTA KÄSITTELEVÄ PALAUTE: johannes.pernaa@edumendo.fi

ESIPUHE

Tervetuloa oppimaan keminformatiikkaa.

Tämän oppaan tarkoituksena on antaa lukijalle intensiivinen esittely nopeasti kehittyvään kemiaan tutkimusalaan nimeltään keminformatiikka. Käsittelen oppaassa tutkimusalan historiaa, molekyylien 2D- ja 3D-visualisoimista tietokoneella, kemiallisen ja kemiaan liittyvän biologisen informaation tallentamista tietokantaan ja tietokantojen käyttöä sekä kemiallisen informaation käsittelyä verkossa ja tieteellisissä julkaisuissa. Tarkastelen myös useita keminformatiikan edistyneempiä teemoja, kuten klusterointia ja monimuotoisuutta, QSAR-malleja ja ennustavaa mallintamista, 3D-kohdistusta ja telakointia sekä keminformatiikan ohjelmistokehitystä. Teosta ei ole suunniteltu oppikirjaksi, vaan monipuoliseksi itseopiskeluoppaaksi. Se on suunnattu teollisuudessa ja akateemisessa maailmassa työskenteleville luonnontieteilijöille ja tietojenkäsittelytieteilijöille, jotka tarvitsevat nopean ja joustavan tutustumisen keminformatiikkaan.

Opas on jaettu 12 teemaan, joista jokainen keskittyy tiettyyn keminformatiikan osa-alueeseen. Tavoitteena on esitellä alan tärkeimmät perusasiat, joiden pohjalta lukija voi erikoistua ja syventää osaamistaan tulevaisuudessa luvun 13 koulutusresurssien avulla.

Ensimmäiset kuusi teemaa keskittyvät alan perusteisiin, kuten esimerkiksi 2D-rakenteiden esittämiseen tietokoneella (teema 2). Teemat 7–12 käsittelevät alan edistyneempiä sovelluksia. Jokaiselle teemalle on määritelty oppimistavoitteet, ja teemojen loppuun on laadittu ajatuksia herättäviä tehtäviä. Tehtävien vastauksen löytyvät liitteestä 1. Oppaassa myös hyödynnetään paljon ulkoisia verkkoresursseja, kuten esimerkiksi tieteellisiä artikkeleita, blogitekstejä ja ohjelmistoja.

Minulla on yli 25 vuoden kokemus keminformatiikasta. Työskentelen nykyisin datatieteiden ja informatiikan professorina Indianan yliopistossa Yhdysvalloissa, jossa johdan yhtä harvoista keminformatiikkaan keskittyvistä koulutusohjelmista. Opettamisen lisäksi johdan tutkimusryhmää, joka on erikoistunut laajamittaiseen tiedonlouhintaan sekä kemiallisen ja biologisen tiedon yhdistämiseen.

Olen perustanut Journal of Cheminformatics -lehden yhdessä Chris Steinbeckin kanssa, ja toimin monien muiden keminformatiikan lehtien toimituskunnissa ja aktiivisena arvioijana. Lisäksi vaikutan aktiivisesti useissa keminformatiikan organisaatioissa, kuten esimerkiksi Chemical Structure Association -säätiön luottamusmiehenä ja American Chemical Societyn jäsenenä. Urani aikana olen osallistunut useiden keminformatiikan konferenssien ja symposiumien järjestämiseen.

Kaikki kommentit sekä korjaus- ja parannusehdotukset tämän oppaan kehittämiseksi ovat erittäin tervetulleita. Toivon, että lähetätte ne minulle osoitteeseen davidjwild@gmail.com. Päivitän alkuperäisteosta säännöllisesti, ja kaikki englanninkielisen oppaan ostaneet saavat päivitetyn version ilmaiseksi.

David Wild

http://djwild.info

SISÄLLYS

1 KEMINFORMATIIKAN HISTORIA JA NYKYTILA

Teeman tavoitteena on oppia

- määrittelemään käsite keminformatiikka
- tuntemaan tutkimusalan historiaa sekä asema kemian ja tietojenkäsittelytieteen rajapinnalla
- mitkä tutkimusongelmat ovat jo ratkaistu, mitkä aiheet ovat aktiivisen tutkimuksen kohteena ja mitkä ovat tärkeimmät tulevaisuuden haasteet
- tuntemaan tutkimusalan tärkeimmät tieteelliset lehdet ja verkkosivustot.

Keminformatiikan määrittely

Tietokoneita on hyödynnetty kemiassa pitkään, mutta termi ”keminformatiikka” otettiin käyttöön vasta 1990-luvulla ensimmäisten kemiaa, tietokonepohjaista laskentaa, informaatiotutkimusta ja lääkeainetutkimusta yhdistävien projektien yhteydessä. Monimuotoisten menetelmien ja käyttötarkoitusten vuoksi käsitteellä ”keminformatiikka” (engl. *cheminformatics*) on monta erilaista määritelmää ja kirjoitusasua, kuten esimerkiksi kemoinformatiikka (engl. *chemoinformatics, suosituin Euroopassa*), kemiallinen informatiikka (engl. *chemical informatics*) ja kemi-informatiikka (engl. *chemi-informatics*). Keminformatiikan tunnetuimpia määritelmiä ovat:

> *”Informaatioteknologian ja tiedonhallinnan hyödyntämistä lääkeainetutkimukseen liittyvässä päätöksenteossa – nopeutetaan datan analysointia informaatioksi ja informaation muuttamista tiedoksi.” (Brown, 1998)*

> *”Informaatiotekniikoiden soveltamista kemian ongelmanratkaisussa.”* (Chemoinformatics: A Textbook, Gasteiger & Engel, 2003)

Tässä teoksessa keminformatiikka määritellään seuraavasti:

> *Keminformatiikka on tutkimusala, jossa tutkitaan kaikkea tietokoneilla esitettävää ja prosessoitavaa kemiallista ja kemiaan liittyvää biologista informaatiota.*

Käytettävä määritelmä valitaan tarpeiden ja tavoitteiden mukaisesti, mutta tekijä suosii yllä mainittua laajaa määritelmää. Määrittelystä on tärkeää ymmärtää, että keminformatiikka on sidoksissa käsitteisiin kuten laskennallinen kemia, molekyylimallinnus ja tietokoneavusteinen lääkesuunnittelu.

- Laskennallisessa kemiassa sovelletaan matemaattisia ja laskennallisia menetelmiä kemian ongelmanratkaisuun. Huomaa, että korkeakouluissa tällä viitataan usein teoreettiseen kemi-

aan, ja termi ”laskennallinen kemia” on käytössä laajemmin teollisuudessa.
- Molekyylimallinnuksessa hyödynnetään 3D-grafiikkaa ja optimointitekniikoita, jotta ymmärrettäisiin yhdisteiden ja proteiinien luonnetta ja käyttäytymistä (sisältää myös materiaalitieteiden sovellukset).
- Tietokoneavusteinen lääkesuunnittelu on tieteenala, jossa käytetään laskennallisia menetelmiä lääkeaineiden löytämisen ja suunnittelun tukena.

Keminformatiikka on sidoksissa myös bioinformatiikkaan, genomiikkaan ja biolääketieteelliseen informatiikkaan.

Keminformatiikan historia

Tietokoneet ja laskennalliset menetelmät ilmestyivät kemian tutkimukseen hyvin aikaisessa vaiheessa. 1950-luvulla alettiin hyödyntää tilastollisia malleja (nykyisin QSAR) ja 1960-luvulla kehitettiin ensimmäiset tietokoneavusteiset kemian visualisoinnit, pääosin aiheesta kiinnostuneiden pioneerien toimesta.

Nykyisessä muodossa tunnetun keminformatiikan pohjatyö 70- ja 80-luvuilla. Alan kehittymistä edisti erityisesti lääketeollisuuden tarve tietokoneavusteiselle lääketutkimukselle.

Keminformatiikan historia ja tulevaisuuden kehittymissuunnat ovat dokumentoitu kolmessa artikkelissa:

- *Chemoinformatics: a history* (Willett, 2011)
- *Chemoinformatics: past, present and future* (Chen, 2006)
- *Grand challenges for cheminformatics* (Wild, 2009).

Näiden lisäksi keminformatiikasta on kirjoitettu muutamia hyviä katsauksia ja perusteoksia:

- *Chemoinformatics—an introduction for computer scientists* - artikkelissa esitellään keminformatiikassa sovellettuja laskennallisia menetelmiä. Artikkeli on laadittu tietojenkäsittelytie-

teilijöille, mutta on helppolukuinen ja soveltuu siten myös muille lukijoille. (Brown, 1998)
- *An introduction to chemoinformatics* -teoksessa käsitellään laajasti keminformatiikan perinteisiä aihealueita. Teoksen ensimmäinen painos ilmestyi vuonna 2003 ja toinen uudistettu painos 2007. (Leach & Gillet, 2007)
- *Chemoinformatics: a textbook* on lyhennetty versio neljäosaisesta keminformatiikan käsikirjasta. Teoksen toimittajat ovat alan varhaisia pioneereja. (Gasteiger & Engel, 2003)

Keminformatiikan perinteisiä sovellusalueita ovat lääkeainetutkimus, kemikaalitietokannat, artikkelien indeksointi ja patenttitietokannat. Uudempia sovelluskohteita ovat reaktiopolkutietokannat, mitta-antureiden kehittäminen, polyfarmakologia, toksikologia jne. Erityisesti julkisen ja avoimen kemiallisen informaation määrä on kasvanut viime aikana nopeasti. Samanaikaisesti keminformatiikan ja lähitutkimusalueiden (esim. bioinformatiikka ja kemiallinen genomiikka) välinen integraatio on lisääntynyt.

Ratkaistuja tutkimusongelmia

Keminformatiikka on laaja ja aktiivinen tutkimuskenttä, mutta joitain aiheita on tutkittu muita intensiivisemmin. Keskeisten kysymysten tutkimiseen on kehitetty laajasti omaksuttuja menetelmiä, joista monet esitellään tässä kirjassa. Yleisesti ajatellaan, että seuraavien ongelmien ratkaisuja pidetään keminformatiikan tutkimusalan menestystarinoina:

Miten 2D- ja 3D-rakenteita esitetään tietokoneella?

Kemiallisten yhdisteiden kuvaamisessa pitää ottaa huomioon muun muassa atomit, sidokset, funktionaaliset ryhmät, nimet ja ominaisuudet. Edellä mainittuja asioita ei ole helppo kuvata tekstipohjaisesti tai millään muullakaan yksinkertaisella tavalla. Keminformatiikassa on ajan saatossa kehitetty hyviä kemiallisten rakenteiden esitystapoja, kuten esimerkiksi lineaarinotaatiot SMILES ja uudempi InChI. Alalle syntyy myös koko ajan uusia tut-

kimuskohteita. Kemiallisten rakenteiden esittäminen verkossa ja massiivisissa julkisissa tietokannoissa tuottavat rakenteiden esittämiselle uusia haasteita, joiden ratkaiseminen vauhdittaa alan tutkimusta.

Miten rakenteita haetaan tietokannoista?

Kemiallisten rakenteiden tiedonhakuun on kehitetty tehokkaita algoritmeja. Niillä voidaan vastata tiedontarpeisiin, kuten ”Mitkä kemialliset yhdisteet tässä yhdisteryhmässä sisältävät tämän alirakenteen (engl. *substructure*)?” tai ”Etsi kaikki kemialliset yhdisteet, joiden rakenne on yhtenevä tämän rakenteen kanssa.”.

Miten kemiallisia rakenteita ja proteiineja visualisoidaan?

Keminformatiikka ja tietokonegrafiikat ovat kehittyneet samanaikaisesti. Kehittämistyö on tuottanut vakiintuneita menetelmiä kemiallisten rakenteiden ja proteiinien visualisoimiselle.

Voidaanko tietokoneilla ennustaa, miten kemikaalit käyttäytyvät koeputkessa tai ihmiskehossa?

Keminformatiikassa on kehitetty laajasti omaksuttuja menetelmiä sille, miten kemiallista rakenneinformaatiota voidaan hyödyntää kemiallisten yhdisteiden ja proteiinikohteiden välisen vuorovaikutuksen biologisten vaikutusten ennustamiseen. Myös esimerkiksi systemaattisten vaikutusten, kuten myrkyllisyyden, selvittämiseen on löydetty uusia menetelmiä. Tämä on nykyisin hyvin aktiivinen tutkimusalue.

Tietenkin on olemassa useita ratkaisemattomia haasteita, joista osaan viitattiin Journal of Cheminformatics -lehden esipuheessa (Wild, 2009), ja joista monia käsitellään myös tässä teoksessa.

Tieteellisiä lehtiä

- Journal of Chemical Information and Modeling: http://pubs.acs.org/journals/jcisd8/index.html
- Journal of Chemical Theory and Computation: http://pubs.acs.org/journals/jctcce
- Journal of Cheminformatics: http://jcheminf.com
- Journal of Computer-Aided Molecular Design: http://www.kluweronline.com/issn/0920-654X
- Journal of Molecular Graphics & Modelling: https://www.sciencedirect.com/journal/journal-of-molecular-graphics-and-modelling
- Journal of Computational Chemistry: https://onlinelibrary.wiley.com/journal/1096987x
- Journal of Medicinal Chemistry: http://pubs.acs.org/journals/jmcmar
- Reviews in Computational Chemistry: https://onlinelibrary.wiley.com/series/6143
- Drug Discovery Today: https://www.journals.elsevier.com/drug-discovery-today
- BMC Bioinformatics: http://www.biomedcentral.com/bmcbioinformatics
- Nature Reviews Drug Discovery: http://www.nature.com/nrd/index.html
- Expert Opinion on Drug Discovery: https://www.tandfonline.com/loi/iedc20

Blogeja, verkkolehtiä ja muita resursseja

- **Scientific Computing World** käsittelee tieteellistä laskentaa eri aloilla. Lehden uutiset liittyvät usein keminformatiikkaan: http://www.scientific-computing.com.
- **Bio-IT World** keskittyy informaatioteknologian hyödyntämiseen biologisessa ja farmaseuttisessa teollisuudessa: http://www.bio-itworld.com.

- **Network Science** verkkolehteä ei enää päivitetä, mutta se sisältää monta tutustumisen arvoista artikkelia: http://www.netsci.org/Science/index.html.
- **CHMINF-L** on kemian kirjastoja ja informaatiotutkimusta käsittelevä sähköpostilista: https://list.indiana.edu/sympa/arc/chminf-l.
- Suositeltavia blogeja:
    - Useful Chemistry -blogi: http://usefulchem.blogspot.com
    - Chem-bla-ics-blogi: http://chem-bla-ics.blogspot.com
    - Noel O'Blog: http://baoilleach.blogspot.com
    - Murray Rustin blogi: http://wwmm.ch.cam.ac.uk/blogs/murrayrust
    - Rajarshi Guhan blogi: http://blog.rguha.net.

Tehtävät

1. Miten bioinformatiikka ja keminformatiikka eroavat toisistaan? Pohdi alojen tutkimuskohteita, historioita ja kulttuureja.
2. Mitkä suuret haasteet on nostettu esille Journal of Cheminformatics -lehden esipuheessa?[1] Oletko niistä samaa mieltä?
3. Lue artikkeli *Systems Chemical Biology* (Oprea ym., 2007).[2] Mitä uusia haasteita tämä asettaa keminformatiikalle?
4. Millaisia keminformatiikan sovelluksia keksit biotieteiden ja lääkeainetutkimuksen ulkopuolelta?
5. Jos voisit suunnitella täydellisen kemiaan räätälöidyn hakukoneen, niin miltä se näyttäisi? Vertaa suunnittelemaasi hakukonetta PubChem-tietokannan ominaisuuksiin ja käyttöliittymään.[3] Oliko PubChemin hakukoneessa jotain samankaltaista, kun sinun suunnitelmassasi?

[1] https://doi.org/10.1186/1758-2946-1-1

[2] https://doi.org/10.1038/nchembio0807-447

[3] http://pubchem.ncbi.nlm.nih.gov

2 KEMIAN 2D-RAKENTEIDEN ESITTÄMINEN TIETOKONEELLA

Teeman tavoitteena on

- oppia kemiallisten rakenteiden historialliset esitystavat
- ymmärtää, miksi rakenteiden esittäminen tietokoneilla on haastavaa
- oppia kuvaamaan yksinkertaisia molekyylejä SMILES- ja InChI-kaavoilla.
- ymmärtää, miten graafiteoriaa hyödynnetään molekyylin rakenneosien välisten suhteiden (esim. atomitaulukko ja sidostaulukko) kuvaamiseen
- tutustua kemiallisten rakenteiden esittämistä koskeviin nyansseihin.

Yhdisteiden historialliset esitystavat

Kun kemia tieteenalana kehittyi satoja vuosia sitten, niin samalla kehittyi useita erilaisia tapoja nimetä kemiallisia yhdisteitä. Jotkin nimeämistavat, kuten esimerkiksi triviaalinimi, toimivat vain yhdisteen yleismaailmallisena tunnisteena. Toiset, kuten molekyylikaava ja systemaattinen nimi, antavat tietoa atomeista ja niiden välisistä sidoksista. Yleisimpiä nimeämistapoja ovat:

- **Triviaalinimi**, esimerkiksi ruokasooda, aspiriini, sitruunahappo jne., on yleiskielessä käytetty ei-systemaattinen nimeämistapa. Vain tunnetuilla ja yleisesti käytössä olevilla kemiallisilla yhdisteillä on triviaalinimi.
- **Molekyylikaava**, esimerkiksi $C_6H_{12}O_6$, on algoritmipohjainen tunniste, joka esittää molekyylin atomit ja niiden lukumäärät, mutta ei anna tietoa yhdisteen rakenteesta (esim. atomien järjestys ja niiden väliset sidokset).
- **Systemaattinen nimi**, esimerkiksi 1,2-dibromi-3-kloori-propaani, kuvaa atomit ja niiden väliset sidokset systemaattisella tavalla, jonka on laatinut IUPAC-järjestö (International Union of Pure and Applied Chemistry).
- **2D-rakennekaava** on kuvaesitys, joka visualisoi molekyylirakenteen atomi- ja sidostasolla.

Nykyään 2D-rakenteet ovat kemian yleiskieltä “lingua franca”. Rakennekaava on suoraviivainen ja joustava kemiallisten rakenteiden esittämistapa, jonka myös muiden alojen tutkijat voivat omaksua helposti. Mutta se mikä on helposti ymmärrettävä muoto ihmisille, ei välttämättä ole sitä tietokoneille. Tietokoneille kuvantunnistus on haastavaa. Tämän vuoksi keminformatiikan pioneerit keskittyivät kehittämään tiedonesitysmalleja, joilla kemiallinen informaatio voitiin esittää tietokoneille ymmärrettävässä muodossa. Merkkijono, joka voidaan helposti konvertoida 1- ja 0-jonoiksi, tunnetaan nimellä linjanotaatio (engl. *line notation*). Esimerkkejä linjanotaatioista ovat seuraavaksi käsiteltävät SMILES- ja InChI-kaavat.

SMILES-kaava

Keminformatiikan varhainen tutkimus pyrki löytämään vastauksia kahteen kysymykseen: 1) Miten rakenneinformaatiota siirretään ihmisiltä tekstipohjaisille tietokoneille? 2) Miten molekyylin atomit ja sidokset esitetään sen jälkeen, kun tieto on tallennettu tietokoneelle? Ensimmäisen kysymyksen ratkaisivat linjanotaatiot, jotka mahdollistivat 2D-rakenteiden kuvaamisen tekstijonoina. Vanhimpia esimerkkejä linjanotaatioista ovat Wiswesser Line Notation (WLN)[4] (Dalke, 2003) ja Beilsteinin ROSDAL, jota käytetään jonkin verran vielä nykyäänkin. Varhaisia tutkimuksia suoritettiin myös siten, että lineaarimerkintöjä käytettiin rakenteiden indeksointiin (sisältäen Lawsonin numeron (engl. *Lawson Number*[5]) (Apodaca, 2010).

Lineaariset esitystavat ovat erittäin käyttökelpoisia vielä nykyäänkin. Niiden käytettävyys ei perustu yksistään siihen, että tietokoneet voivat työstää vain tekstiä, mutta koska teksti on yhä informaation tehokkain tallennus- ja viestintämuoto. Rakenteen lineaarinen esitysmuoto voidaan tallentaa taulukon soluun tai tietokannan tekstikenttään. Käytetyimmät linjanotaatiot ovat SMILES (Simplified Molecular Input Line Entry System) ja InChI.

SMILES-kaavassa atomit esitetään kirjaimilla. Kuusi yleisintä orgaanisten yhdisteiden atomia esitetään suoraan kemiallisella merkillä (C, N, S, O, P, H). Muut atomit ympäröidään hakasulkeilla, kuten esimerkiksi [Ag] tai [Cu]. Tekstijonon vierekkäisten symbolien oletetaan sitoutuvan toisiinsa yksinkertaisella sidoksella. Tämän säännön mukaan alkaanit voidaan esittää SMILES-tekstijonona seuraavasti:

C metaani
CC etaani
CCC propaani
CCCC butaani.

[4] http://www.dalkescientific.com/writings/diary/archive/2003/10/15/WLN.html

[5] http://depth-first.com/articles/2010/09/28/a-brief-introduction-to-lawson-numbers

Huomaa, että vedyt generoidaan epäsuorasti atomin sidoskertaluvusta, joten niitä ei tarvitse kuvata SMILES-kaavassa. Tyhjät sidospaikat täytetään vedyillä automaattisesti, kaksoissidos kuvataan =-merkillä ja kolmoissidos #-merkillä. Nyt osaat kuvata jo suuren joukon rakenteita, kuten esimerkiksi:

C=C	eteeni
C=CC	propeeni (voidaan esittää myös muodossa CC=C)
[C-]#N	syanidi-ioni.

Viimeisessä esimerkissä syanidi-ionin hiiliatomille on lisätty varaus. Tämä vaatii hakasulkeiden lisäämisen hiiliatomin ja tavuviivan ympärille, mikä erottaa ne vierekkäisestä atomista. Huomaa myös, että propeeni voidaan esittää kahdella eri tavalla. SMILES kuvaa kemiallisen rakenteen, mutta ei yksilöllisesti, joten sitä ei voi käyttää kemiallisena tunnisteena. Tämä johtuu osittain kanonisoinnista (katso Morgan algoritmin kuvaus (Evans, 2011).

Molekyylin haaroittuneisuus ilmaistaan sulkeiden ( ) avulla. Haarat voivat myös olla sisäkkäisiä. Alla muutama esimerkki haaroittuneista rakenteista:

CC(C)C	isobutaani
CNCCC(=O)OH	3-(metyyliamino)propaanihappo
CC(CC(=O)OH)CCN	5-amino-3-metyylipentaanihappo.

Yksinkertaiset SMILES-rengasrakenteet voidaan sulkea numeroilla (1–9, tai 10–99 hakasulkeissa). Eli sykloheksaanin SMILES-kaava olisi C1CCCCC1, jossa kaksi 1:stä ilmaisevat renkaan sulkevien sidosten paikat. Renkaan aromaattisuutta ei ilmaista sidosten avulla vaan atomeilla. Toisin sanoen, atomi koetaan aromaattiseksi eikä sidos. Atomi määritellään aromaattiseksi pienellä kirjaimella. Täten aromaattisten ja ei-aromaattisten rengasrakenteiden ero kuvataan seuraavasti:

C1CCCCC1	sykloheksaani
c1ccccc1	bentseeni.

Huomaa, että renkaan sulkeva numeromerkintä ilmaisee pelkästään kahden atomin välistä epätavallista sidosta. Se ei määrittele suoraan renkaan sulkeutumista, vaikka se on merkinnän pääasiallinen tehtävä. SMILES-kaavaan voidaan tehdä myös eksoottisia rakennemuunnelmia esimerkiksi pisteen "." avulla. Piste katkaisee vierekkäisten atomien välisen sidoksen.

C1C.CC1	butaani
C1CC1.C2CC2	sykloheksaani

Tätä merkintätekniikkaa hyödynnetään esimerkiksi ohjelmistoissa, jotka yhdistelevät molekyylejä SMILES-kaavoiksi.

SMILES-tekniikasta löytää lisätietoja Daylight SMILES -sivustolta.[6] Sivustolla on SMILESin teoriaa, käyttöohjeet, tutoriaaleja ja apuohjelmistoja, joiden avulla voit kuvailla omat rakenteesi SMILES-kaavoina. SMILESistä on olemassa myös avoimen lähdekoodin versio, jota ylläpitää ja kehittää OpenSMILES-yhteisö.[7]

InChI-kaava

InChI (International Chemical Identifier) kehitettiin 2000-luvun alussa IUPACin and NISTin toimesta. Tavoitteena oli kehittää standardi kemiallisten rakenteiden tunnistamiseen tietokoneilla ja tarjota kanoninen ohjelmistopaketti InChI-kaavojen tuottamiseen. InChI suunniteltiin välttämään SMILES-kaavan ongelmaa – erilaiset toteutustavat voivat tulkita saman SMILESin eri tavoin. InChI ottaa SMILESia paremmin huomioon myös kemiallisia nyansseja, kuten stereokemiaa ja tautomeriaa.

InChI kuvailee kemiallisen rakenteen tasoittain – jokaisen tason avulla kuvaillaan jokin yhdisteen ominaisuus. Tasot erotetaan kenoviivalla "/", ja vain päätaso on pakollinen. Yleisimmin käytetyt tasot ovat:

[6] http://www.daylight.com/smiles

[7] http://www.opensmiles.org

- Päätaso: kemiallinen kaava, sidokset (c-etuliite), vedyt (h-etuliite).
- Varaustaso: positiiviset varaukset (p) ja negatiiviset varaukset (q).
- Stereokemiallinen taso: kaksoissidokset (b), tetraedrinen stereokemia (t ja m) ja stereokemiatyyppi (s).

Tällä hetkellä kaikki InChI-merkinnät sisältävät etuliitteen "InChI=". Etuliitettä seuraa osoitin ("1/" tai "1S/"), joka ilmaisee, onko kyseessä standardien mukainen InChI-merkintä (esim. ohjelmiston mukaiset standardiasetukset) vai poikkeaako se standardiasetuksesta (engl. *non-standard*). Alla on esitetty muutamia pelkästään päätasosta koostuvia esimerkkejä:

InChI=1S/CH4/h1H4	metaani
InChI=1S/C3H6/c1-3-2/h3H,1H2,2H3	propeeni
InChI=1S/C6H12/c1-2-4-6-5-3-1/h1-6H2	sykloheksaani
InChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H	bentseeni
InChI=1S/C4H9NO2/c1-5-3-2-4(6)7/h5H,2-3H2,1H3,(H,6,7)	3-metyyli-aminopropaanihappo.

Kemiallinen kaava on riittävän suoraviivainen, mutta sitoutuminen ja vetyjen kuvaaminen vaativat hieman lisäselvitystä. Sidosjakso kuvaa pää- ja sivuketjut – esimerkiksi yläpuolella esitetyssä propeenin InChI-kaavassa ensimmäinen atomi (1) on sitoutunut kolmanteen atomiin (3), joka taas on sitoutunut toiseen atomiin (2). Viimeisessä esimerkissä on haaroittunut rakenne, jonka haaroittuneisuus esitetään sulkeilla. Tässä mielessä sidosrakenne kuvataan samalla tavalla kuten SMILES-kaavassa, mutta atomit kuvataan atominumeroilla kemiallisen merkin sijaan. Atomien numerot asetetaan kemiallisessa kaavassa esitetyn järjestyksen mukaisesti. Sidoskertalukua ei määritetä sidosjakson perusteella, vaan vetyjen paikat määräävän vetyjakson avulla. Esimerkiksi 2-4H2 tarkoittaa, että atomeihin 2–4 on sitoutunut kaksi vetyatomia; 1-3,6,7H tarkoittaa, että atomeihin 1–3, 6 ja 7 on sitoutunut yksi vetyatomi. "Liikkuvat vetyatomit" (engl. *mobile hydrogens*) merkitään sulkeilla – ne voivat liikkua eri atomien välillä.

Myös yhdisteelle voidaan generoida InChI-avain. Se on erillinen InChI-lineaarimerkinnästä, ja sitä käytetään erityisesti internetin tiedonhakukoneille sopivien yhdisteiden tunnistamiseen. Se on ASCII-merkistöön pohjautuva InChI-lineaarimerkinnän hajautus, mutta sen pituus on kiinteä. Lisäksi siinä käytetään vain merkkejä, joita ei yleensä pidetä erottimina. InChI-lineaarimerkinnän eri osat esitetään erillisinä hajautusjaksoina väliviivoin eroteltuna, joten verkkohaussa ensimmäinen osa palauttaa toisia hakuun liittyviä isomeerejä jne. Ohessa on muutamia PubChemistä poimittuja esimerkkejä:

VDIPNVCWMXZNFY-UHFFFAOYSA-N	3-metyyli-aminopropaanihappo
VNWKTOKETHGBQD-UHFFFAOYSA-N	metaani
QQONPFPTGQHPMA-UHFFFAOYSA-N	propeeni.

InChI-merkinnän eri tasoista ja niiden sisällöstä löydät lisätietoa InChI:n virallisilta kotisivuilta.[8] Sivuilta löydät täydellisen dokumentaation, InChI:n käyttöoppaan, ohjelmatiedostoja ja -koodia sekä InChI FAQ:n.[9] Merkintätavan kehittämistä tukee InChI-säätiö.[10]

Graafiteorian sisäinen esitys

Graafiteoria on matemaattinen työkalu, jota käytetään verkkojen mallintamiseen. Verkko koostuu solmuista (engl. *node*) ja niiden välisistä linkeistä, joita kutsutaan väleiksi (engl. *edge*). Miten verkkoja sovelletaan kemiallisten rakenteiden mallintamiseen? No, jos atomien ajatellaan olevan solmuja ja sidosten välejä, niin kemiallinen rakenne on matemaattinen verkko.

Verkkojen käsittelyyn on kehitetty useita geneerisiä algoritmeja. Esimerkiksi verkkojen isomorfia -algoritmin avulla voidaan

[8] http://www.iupac.org/inchi

[9] https://www.inchi-trust.org/faq

[10] http://www.inchi-trust.org

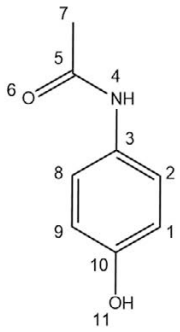
verrata kahta kemiallista rakennetta toisiinsa ja selvittää, vastaavatko ne toisiaan. Aliverkon isomorfia -algoritmilla voidaan selvittää, sisältääkö kemiallinen rakenne analysoitavan alarakenteen. Vakiintuneet algoritmit soveltuvat lähes sellaisenaan edellä mainittujen kysymysten ratkaisemiseen. Niiden käytössä täytyy ottaa huomioon kemiallisten rakenteiden esittämistapojen useat vivahteet, jotka kuvataan myöhemmin tässä luvussa.

2D-rakenteiden sisäinen kuvaus hyödyntää pääosin verkkojen mallintamisen standardimetodeja, sisältäen muutaman kemialle tärkeän lisäyksen. Jokaiselle atomille (tai solmulle) on asetettu yksilöllinen numero, joka tallennetaan atomin kemiallisen merkin yhteyteen. Minimissään kuvaus sisältää tiedon alkuaineen kemiallisesta merkistä (esim. C, N, S), mutta joskus käytetään edistyneempää kuvausta, mikä mahdollistaa esimerkiksi hybridisaatioiden ja muiden ominaisuuksien kuvaamisen. Tiedot voidaan esittää taulukkona, jota kutsutaan toisinaan *atomihakutaulukoksi* (engl. *atom lookup table*) (ks. kuva 1). Keminformatiikassa taulukkoa, joka esittää atomeja yhdistävät sidokset, kutsutaan yleisesti liitostaulukoksi (tai yhteystaulukoksi) (engl. *connection table*), vaikka se on muodoltaan matemaattinen vierusmatriisi (engl. *adjacency matrix*).

Liitostaulukossa jokaiselle atomille on varattu sarake ja rivi. Atomeja yhdistävät sidokset ilmaistaan sarakkeiden ja rivien risteämispisteissä numeroilla (1 tai suurempi). Jos atomien välillä ei ole sidosta, niin risteyskohtaan merkitään nolla. Saman atomin sarakkeen ja rivin risteyskohdan merkitsemistavalle ei ole päätetty yhtä oikeaa tapaa. Kuvan 1 esimerkissä saman atomin risteyskohtaan on merkattu nolla. Yleensä sidoskertaluku ilmaistaan asettamalla risteyskohdassa numeroin seuraavasti: 1=yksinkertainen sidos, 2=kaksoissidos, 3=kolmoissidos. Joskus luvulla 4 kuvataan aromaattista sidosta. Koska liitostaulukko on toisteinen (engl. *redundant*) (se ilmaisee samanaikaisesta atomin *a* yhteyden atomiin *b* ja toisinpäin), liitostaulukon ei-toisteista (engl. *non-redundant*) muotoa voidaan käyttää, jos informaatio halutaan tallentaa vain kerran. Kuvassa 1 on esitetty atomihakutaulukko ja

liitostaulukko asetaminofeenille (Tylenoli, Parasetamoli), jossa taulukon ei-toisteinen osa on vahvennettu.

Atomi-numero	Atomi-tyyppi
1	C
2	C
3	C
4	N
5	C
6	O
7	C
8	C
9	C
10	C
11	O



Liitostaulukko											
	1	2	3	4	5	6	7	8	9	10	11
1	0	1	0	0	0	0	0	0	0	2	0
2	**1**	0	2	0	0	0	0	0	0	0	0
3	**0**	**2**	0	1	0	0	0	1	0	0	0
4	**0**	**0**	**1**	0	1	0	0	0	0	0	0
5	**0**	**0**	**0**	**1**	0	2	1	0	0	0	0
6	**0**	**0**	**0**	**0**	**2**	0	0	0	0	0	0
7	**0**	**0**	**0**	**0**	**1**	**0**	0	0	0	0	0
8	**0**	**0**	**1**	**0**	**0**	**0**	**0**	0	2	0	0
9	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	0	1	0
10	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	0	1
11	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	0

Kuva 1. Vasemmalle on esimerkki atomihakutaulukosta. Oikealla on vastaava liitostaulukko, jonka ei-toisteinen osa on esitetty vahvennettuna.

On tärkeää, että sama molekyyli numeroidaan samalla tavalla joka kerta. Tarvitaan algoritmi, joka määrittää numerot johdonmukaisesti sääntöjen perusteella. Onneksi tämä voidaan tehdä Morgan algoritmilla (Evans, 2011). Algoritmissa jokaiselle atomille asetetaan sitoutumisarvo (engl. *connectivity value*) sen mukaan, miten moneen atomiin se on sitoutunut. Tämä arvo korvataan iteratiivisesti naapuriatomien arvojen summalla niin kauan, kunnes eri arvojen lukumäärä on maksimoitu. Seuraavaksi atomit numeroidaan sitoutumisarvojen mukaisesti laskevaan järjestykseen. Arvojen ollessa yhtä suuria, asetetaan järjestys muiden ominaisuuksien avulla (esim. atominumero, sidoskertaluku jne.).

Säännönmukainen toiminta on kanonisten esitystapojen keskeinen perusta. Morgan algoritmia käytetään erityisesti linjanotaatioiden kanonisointiin, jonka avulla varmistetaan molekyylin yhteystaulukoiden samankaltainen numerointi jokaisella kerralla.

Tiedostotyypit

Linjanotaatiot eivät ole ainoa tapa kommunikoida rakenteiden kanssa. Myös tiedostopohjaiset muodot, kuten MDL:n MOL-tiedosto[11], MOL-tiedoston muunnos SD-tiedosto ja Chemical Markup Language -tiedosto (CML)[12] (CML on muunnelma XML-tiedostosta), ovat suosittuja. Mainitut tiedostomuodot ovat niin sanottuja dump-tiedostoja. Ne sisältävät minimissään atomihaku-taulukon ja yhteystaulukon, mutta niihin voidaan tarpeen mukaan liittää myös muita tietokenttiä. Lisätietojen lisäämismahdollisuus on tiedostojen vahvuus, minkä vuoksi ne soveltuvat monen tyyppisiin tarpeisiin.

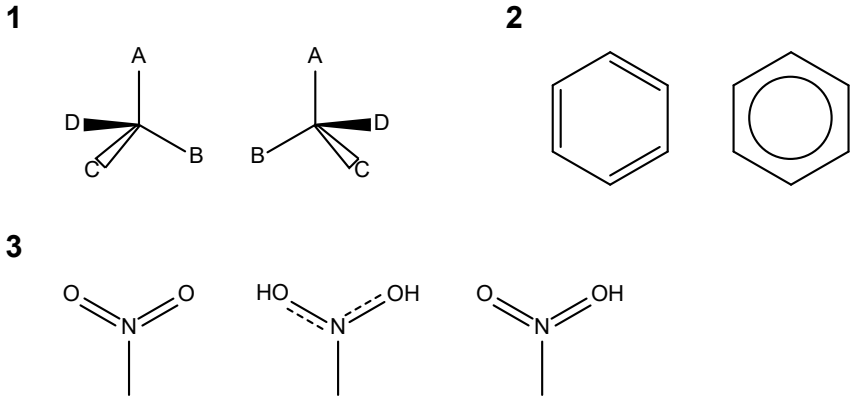
Rakennekuvauksen nyanssit

Tähän mennessä on esitelty joitain yksinkertaisia tapoja, miten 2D-rakenteita esitetään ja miten niiden kanssa kommunikoidaan. Kemiallisten rakenteiden esittämisessä on lisäksi muutamia nyansseja, jotka monimutkaistavat asioita – erityisesti stereokemia, aromaattisuus ja tautomeria (ks. kuva 2).

SMILES-kaava ei sisällä tietoa stereokemiasta, mutta se voidaan tallentaa InChI-kaavaan tai isomeeriseen SMILES-kaavaan (engl. *Isomeric SMILES*). Normaalisti 2D-rakenteissa stereokemia kuvataan kiiloilla, jotka osoittavat, nouseeko sidos ylöspäin vai alaspäin (ks. kuvan 2 visualisointi 1).

[11] MDL-tiedoston (2005) määrittelydokumentti: http://c4.cabrillo.edu/404/ctfile.pdf

[12] http://www.xml-cml.org



Kuva 2. Esimerkkejä SMILES-kaavoille haastavista rakenteista (1=stereokemia, 2=aromaattisuus ja 3=tautomeria).

Vaikka stereokemian merkitseminen esitykseen on mahdollista, niiden esittämisessä on seuraava haaste: joissain tilanteissa stereoisomeerit halutaan erottaa ja välistä niitä on tarpeen käsitellä samana rakenteena. Esimerkkinä voidaan tarkastella talidomidia, jolla on kaksi enantiomeeriä. Välistä yhdiste halutaan tallentaa tietokantaan itsenäisenä yhdisteenä siten, että enantiomeerejä ei erotella, mutta toisinaan ne taas halutaan erottaa, koska enantiomeereillä on toisistaan eroavia biologisia ominaisuuksia.

InChI-kaavassa tämä on ratkaistu siten, että stereokemia esitetään erillisenä osana notaatiota. Vaikka stereokemiaa ei merkittäisi kaavaan, on yhdisteen InChl-kuvaus silti oikeaoppinen – se ei vain sisällä tietoa yhdisteen mahdollisesta stereokemiasta. Aromaattisuuden haasteena on elektronien taipumus delokalisoitua tasaisesti rengasrakenteessa, mikä vaikeuttaa yksöis- ja kaksoissidosten erottamista toisistaan. Rengasrakenteiden aromaattiseksi ja ei-aromaattiseksi kategorisointiin on useita lähestymistapoja, kuten esimerkiksi Hückelin 4n+2-sääntö. Mahdollisuus rengasrakenteiden monimuotoiseen esittämiseen (yksinkertaisen sidoksen ja kaksoissidoksen vuorottelu, aromaattiset atomit tai sidokset) voi johtaa sekaannuksiin ja virheisiin molekyylien vertailussa. Esimerkiksi Kekulén ja jonkin muunlaisen bentseenirakenteen esityksen tunnistaminen samaksi rakenteeksi vaatii siihen pystyvän

algoritmin. Huomaa, että aromaattisuutta voidaan käsitellä sekä esitystasolla (aromaattisten atomien merkitseminen, aromaattiset sidokset tai yksöis- ja kaksoissidoksen vuorottelu) että algoritmitasolla (aromaattisuuden määrittely rakenneanalyysillä).

Myös tautomerian (atomin tai atomiryhmän siirtyminen paikasta toiseen molekyylin sisällä) käsittely on haastavaa. Esimerkki tautomeriasta on esitetty kuvan 2 alaosassa, jossa on visualisoitu nitroryhmän kolme erilaista muotoa. Ensimmäinen rakenne (pentavalentti muoto (engl. *pentavalent form*) ei ole kemiallisesti validi ilman varausta, mutta sitä käytetään yleisesti kuvauksessa. Toisessa käytetään aromaattisia sidostyyppejä osoittamaan, että elektronit delokalisoituvat happiatomien välillä. Kolmannessa on selkeästi ilmaistu, että yksi happi on sitoutunut kaksoissidoksella ja toinen yksinkertaisella sidoksella.

Kaikkia muotoja pidetään kelvollisina, mutta kontekstin mukaan voi olla tarpeellista esittää joko yksi tarkka kuvaus (esim. vain muoto 3 voi esiintyä tietyssä pH:ssa) tai olla epämääräinen muodon suhteen. InChI:ssa useat tautomeriset muodot voidaan kuvata samanaikaisesti yhdellä rivillä sisällyttämällä tautomeriakuvaukset kaavaan.

Reaktioiden ja geneeristen rakenteiden esittäminen

Reaktioiden rakenteellisissa esityksissä on yksilöitävä vain tuotteiden ja reagenssien järjestys ja mahdollisesti, mikä lähtöaineineen atomi on tuotteen vastaava atomi. Muu informaatio, kuten stoikiometria ja saanto tallennetaan yleensä erikseen. Reaktio-SMILES on SMILES-merkinnän yläjoukko, jossa on symbolit nuolille ja reaktion eri komponenteille. Reagenssit, katalyytit ja tuotteet erotetaan toisistaan >-merkillä, ja yksittäiset reagenssit, katalyytit ja tuotteet erotetaan toisistaan pisteellä. Esimerkiksi metaanin palaminen hiilidioksidiksi ja vedeksi voitaisiin esittää

C.OO>>C(O)O.O

Huomaa, että koska ei ole katalyyttiä, ovat molemmat >-merkit peräkkäin. Myös lähtöaineiden atomien yhdistäminen tuotteiden atomeihin on mahdollista. Tämä toteutetaan atomileimoilla (esimerkki otettu Daylight-tutoriaalista):

[CH2:1]=[CH:2][CH:3]=[CH:4][CH2:5][H:6]>>
[H:6][CH2:1][CH:2]=[CH:3][CH:4]=[CH2:5]

Toista kieltä, SMIRKS-kaavaa, voidaan käyttää moniselitteisyyden sallimiseksi reaktioissa. Tämä mahdollistaa muunnoksien esittämisen. Esimerkiksi voidaan esittää alirakenteen tai funktionaalisen ryhmän muutos ilman koko reaktion esittämistä. Seuraava esimerkki on otettu Daylight-tutoriaalista:

[*:1][N:2](=[O:3])=[O:4]>>[*:1][N+:2](=[O:3])[O-:4]

Teknisesti SMIRKS on SMILES- ja SMARTS-hybridi (kyselykieli, joka käsitellään seuraavassa teemassa). Se on erityisen hyödyllinen geneeristen sovellettavien reaktioiden esittämiseen. Erinomainen esimerkki sen käytöstä on DrugGuru tool -työkalussa (Stewart ym., 2006).

Tutustu reaktio-SMILESiin ja SMIRKSiin syvemmin Daylighting Reaction SMILES and SMIRKS -tutoriaalin[13] ja Daylight Theory -ohjeen SMIRKS-luvun kautta[14].

Geneeristen rakenteiden esittäminen on eri haaste, mutta liittyy hieman aiheeseen. Kemiallisten rakenteiden geneeriset muodot esiteltiin todennäköisesti ensimmäisen kerran vuonna 1924 osana patenttia Eugene Markushin toimesta (ennen sitä patentit kohdistuivat tiettyihin rakenteisiin). Siksi termiä "Markush-rakenteet" alettiin käyttää 2D-esityksistä, jotka kuvaavat useampaa kuin yhtä tiettyä rakennetta, esimerkiksi luetteloimalla vaihto-

[13] Daylight Theory: SMIRKS - A Reaction Transform Language: https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html

[14] Reaktio-SMILES ja SMIRKS: https://www.daylight.com/meetings/summerschool01/course/basics/smirks.html

ehtoiset ryhmät tietyillä molekyylien kohdilla tai määrittelemällä monimuotoisuus kiinnityskohdassa. Geneeristen rakenteiden esittäminen on vaikeaa, koska Markush-rakenne voi edustaa rajoittamatonta määrää yhdisteitä (esim. "aryyliryhmä").

Tätä ongelmaa on käsitelty kuvaamalla geneerisiä rakenteita tekstipohjaisilla kielillä, kuten esimerkiksi GENSAL-kielellä, ja käyttämällä laajennettuja yhteystaulukoita. Niitä käytetään laajasti patenttihakujärjestelmissä. Lisätietoa geneerisistä rakenteista löytyy artikkelista *The Sheffield Generic Structures Project – a Retropsective Review* (Lynch & Holliday, 1996)[15].

Tehtävät

6. Esitä ibuprofeenille SMILES- ja InChI-kaavat.
7. Suorita Google-haku termille **WTDRDQBEARUVNC-LURJTMIESA-N**. Minkä lääkkeen InChI-avain on kyseessä? Seuraavaksi hae **WTDRDQBEARUVNC-ZCFIWIBFSA-N**. Mihin tarkoitukseen tämä InChI-avain on? Huomaa, että ensimmäinen osa on sama. Miten yhdisteet eroavat? Lopuksi hae vain ensimmäisellä osalla **WTDRDQBEARUVNC.** Löydätkö molemmat?
8. Kirjoita Daylight Depict -työkaluun seuraava SMILES: C1=CC=C2C(=C1)NC=N2. Mikä yhdiste on kyseessä? Huomaa, että molemmat renkaat on esitetty aromaattisessa muodossa, mutta SMILES syötettiin Kekulé-muodossa. Miksi aromaattinen muoto esitettiin?
9. Kokeile vaihtoehtoisia muotoja: c1ccc2c(c1)NC=N2 ja c1ccc2c(c1)nc=n2. Miksi jälkimmäinen muoto hylätään?

[15] http://pubs.acs.org/doi/abs/10.1021/ci950173l

3 2D-RAKENTEIDEN KARAKTERISOINTI DESKRIPTOREILLA JA SORMENJÄLJILLÄ

Teeman tavoitteena on

- tutustua kemiallisten rakenteiden kuvaamiseen soveltuviin deskriptoreihin
- pystyä erottamaan fragmentaariset (suom. *epäyhteinäinen;* engl. *fragmental)*, fysikaaliskemialliset (engl. *physicochemical)* deskriptorit ja topologiset indeksit toisistaan
- ymmärtää käsitteet rakenteellinen avain (engl. *structural key*) ja sormenjälki
- oppia yleisimmät metodit, jotka soveltuvat sormenjälkien samankaltaisuuksien laskentaan.
- tutustua kemiallisten rakenteiden esittämistä koskeviin nyansseihin.

Johdanto

Tähän saakka käsitellyt metodit soveltuvat pääosin kemiallisten rakenteiden esittämiseen ja tunnistamiseen. Lisäksi on metodeja rakenteiden ominaisuuksien kuvaamiseen, joita kutsutaan deskriptoreiksi. Seuraavaksi tarkastellaan, miten deskriptoreita tuotetaan ja miten niitä käytetään yhdisteiden karakterisoinnin ”sormenjälkinä”. Deskriptorit soveltuvat useisiin käyttötarkoituksiin, mutta erityisesti ennustavaan mallintamiseen ja rakenteiden samankaltaisuuksien analysoimiseen. 2D-rakenteesta voidaan laskea deskriptoreita, kuten

- yksinkertaisten ominaisuuksien lukumääriä (esim. pyörivien sidosten lukumäärä tai molekyylipaino)
- fragmentaariset deskriptorit, jotka osoittavat todellisen tai geneerisen alirakenteen läsnäolon, poissaolon tai lukumäärän
- fysikaaliskemialliset ominaisuudet
- topologiset indeksit, kuten esimerkiksi haarautumisindeksi (engl. *branching index)* ja khiin molekulaarinen sitoutumisindeksi (engl. *chi molecular connectivity indice*).

Laajemman listan löydät Molconn-Z Methods -sivustolta.[16]

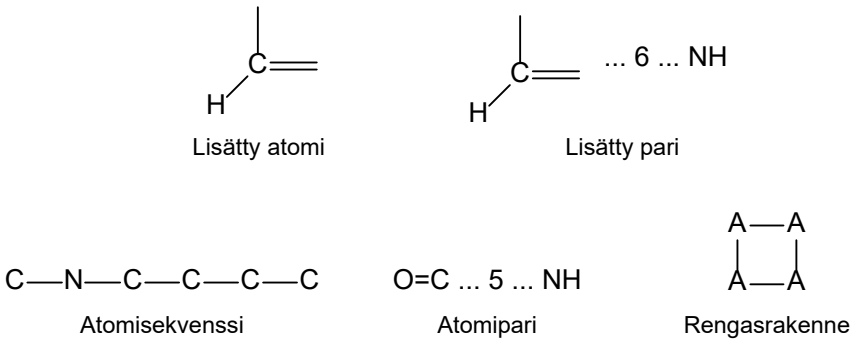
Fragmentaariset deskriptorit

Fragmentaariset deskriptorit kuvaavat yhtä atomia laajempia 2D-rakenteen ominaisuuksia. Tällaisia ominaisuuksia ovat tietyt alirakenteet (esim. nitro- tai karboksyylihapporyhmä), yksinkertaiset alirakenteet tai monimutkaisemmat rakenteet. Fragmentaariset deskriptorit voivat pohjautua sääntöihin (generoitu sääntöjen avulla molekyyliaineistosta) tai sanakirjaan (sanakirjassa määritetyt sattumanvaraiset alirakenteet).

Kuvassa 3 on esimerkkejä fragmentaarisista deskriptoreista, kuten lisätyt atomit (atomit ja niitä ympäröivät sidokset) (engl.

[16] http://www.edusoft-lc.com/molconn/manuals/400/methodex.html

augmented atom), atomisekvenssit (kaikkien tietyn alueen sitoutuneiden atomien polut) (engl. *atom sequence*), atomiparit (engl. *atom pair*) ja lisätyt parit (atomit ja lisätyt atomit, joiden välissä on tietty määrä sidoksia) (engl. *augmented pair*) ja rengasrakenteet (rengassysteemien polut) (engl. *ring composition*). Sääntötyypin mukaan, useilla algoritmeilla voidaan tunnistaa molekyylissä olevia fragmentteja. Sanakirjaan pohjautuvat fragmentaariset deskriptorit määrittävät kiinnostuksen kohteena olevista alirakenteista listan tai sanaston. Ne kirjataan usein esimerkiksi SMARTS-kaavoina. Kummassakin tapauksessa tietty alajoukko valideja deskriptoreita on läsnä tietyssä molekyylissä, jotka olisivat kyseisen molekyylin deskriptorit.



Kuva 3. Esimerkkejä fragmentaarisista deskriptoreista.[17]

Fysikaaliskemialliset ominaisuudet

Molekyylin fysikaaliset ja kemialliset ominaisuudet voidaan määrittää kokeellisesti tai arvioida algoritmeilla 2D-rakenteen pohjalta. Ominaisuudet esitetään usein numeroilla. Yleisiä fysikaaliskemiallisia deskriptoreita ovat esimerkiksi LogP (kertoo yhdisteen öljyisyyden, joka vaikuttaa sen kulkeutumiseen kehossa), mole-

[17] Suomentajan kommentti: Rengasrakenteen mallia on muokattu alkuperäisestä selkeämmäksi. ”A” viittaa atomiin.

kyylipaino, vetysidosten luovuttajien ja vastaanottajien määrä ja polaarinen pinta-ala. Näiden deskriptorien avulla on tehty joitain mielenkiintoisia analyysejä. Esimerkiksi Lipinskin viiden sääntö[18] asettaa kriteerit tietyille deskriptoreille markkinoitujen lääkkeiden tilastollisen analyysin perusteella.

Topologiset indeksit

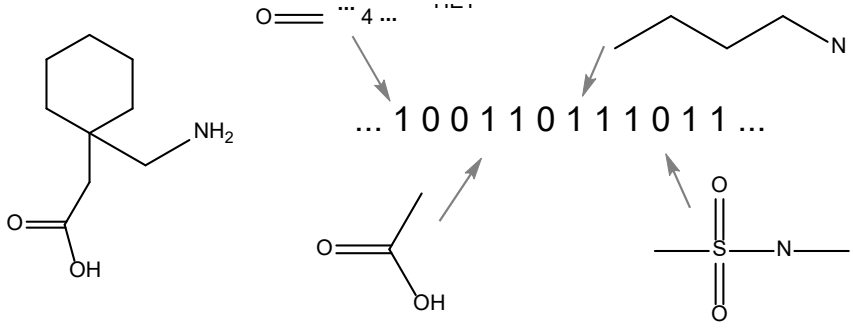
Topologiset indeksit ovat yksiarvoisia deskriptoreita, jotka kuvaavat jotain kemiallisen rakenteen kuvaajan ominaisuutta (siksi nimi ”topologinen”). Eräs ensimmäisistä oli Wiener-indeksi, joka on yksinkertaisesti 0,5 x kaikkien atomiparien välisten sidosten lukumäärän summa. Myöhemmin Wiener-indeksiin sisällytettiin molekulaariset yhteysindeksit, Randic-haarautuvuusindeksit sekä Kier- ja Hall khiin -molekulaariset sitoutumisindeksit. Lisätietoja näistä löytyy jo mainitusta Molconn-Z-menetelmien käsikirjasta.[19]

Deskriptoreista sormenjäljiksi

Deskriptoreista on helppo koota merkkijono (string), joka karakterisoi yhdisteen. Deskriptorit voivat olla binäärisiä (1 osoittaa ominaisuuden olemassaoloa ja 0 sen puuttumista) (ks. kuva 4), numeerisia (kokonaislukuja, reaalilukuja jne.) tai kategorisia. Keminformatiikassa näitä deskriptorijonoja kutsutaan sormenjäljiksi. Binaarideskriptorit ovat erityisen hyödyllisiä, koska binäärisarjojen käsittelyyn on olemassa tehokkaita algoritmeja.

[18] http://en.wikipedia.org/wiki/Lipinski's_Rule_of_Five

[19] http://www.edusoft-lc.com/molconn/manuals/400/methodex.html



Kuva 4. Esimerkki binäärisestä sormenjäljestä.

Yksinkertaisimmassa tapauksessa deskriptoreja ja niille varattuja paikkoja sormenjäljessä on 1:1-suhteessa. Esimerkiksi yleinen käyttötarve on saada fragmentaarisista 2D-deskriptoreista binaarinen sormenjälki, joissa yksi bittijonon bittipaikka yhdistetään yhteen sanakirjan tietueeseen. Bittiarvo (1,0) määrittelee kyseisen ominaisuuden esiintymisen tai puuttumisen.

Tällaista sormenjälkeä kutsutaan joskus rakenteelliseksi avaimeksi, joista keminformatiikan tunnetuin esimerkki on MDL 166-bittinen rakenneavain (tunnetaan joskus MACCS- tai ISIS-avaimena). MDL 166 määrittelee 166 fragmenttia, joita pidetään tärkeinä lääkeainekemiassa.

Vaihtoehtoinen strategia sormenjälkien tuottamiseksi on käyttää sääntöpohjaisia fragmentaarisia deskriptoreita, jolloin molekyylin tai molekyylijoukon deskriptorit luodaan lennossa. Esimerkkejä yleisistä säännöistä ovat

- kaikki atomisekvenssit atomien 2–7 väliltä
- kaikki lisätyt atomit
- pyöreät alirakenteet.

Ilman sanakirjaa ei ole johdonmukaista tapaa kartoittaa deskriptoreita sormenjäljen bitteihin. Myös generoitujen fragmenttien lukumäärä voi olla valtava (esim. 100 000 vain atomien 2–7 sekvensseille C, N, S, O, P ilman, että otetaan huomioon sidostyyppejä tai yleistyksiä). Jos jokaiselle deskriptorille luotaisiin

bittipaikka, olisivat sormenjäljet todella suuria ja harvoja. Tämän vuoksi yleensä deskriptorit kartoitetaan kiinteälle määrälle bittejä (esim. 1 024) hajautusalgoritmien avulla. Tällaisia sormenjälkiä kutsutaan hajautetuiksi sormenjäljiksi.

Binääriset sormenjäljet ja rakenteelliset avaimet ovat saatavilla monesta lähteestä:

- MDL (166-avaimet, saatavilla monessa eri muodossa)
- Scitegic ECFPs (sisällytetty Pipeline Pilot -pakettiin)
- Daylight hashed fingerprints
- BCI fingerprints
- CDK (Chemistry Development Kit)
- Chemaxon.

Ei-binääriset sormenjäljet voidaan luoda käyttämällä fragmenttikuvausten varianttia (esim. numero, joka ilmaisee, kuinka monta kertaa fragmentti esiintyy molekyylissä), kategorisia tai ei-binäärisiä fragmentteja.

Sormenjälkien samankaltaisuuksien mittaaminen

Yleisin tapa mitata kahden sormenjäljen välinen samankaltaisuus on Tanimoto-kerroin. Binaarisen sormenjäljen tapauksessa Tanimoto on identtinen tunnetumman Jaccard-indeksin kanssa. Yleensä tämä on määritelty joukon leikkauspisteeksi, joka jaetaan joukon liitoksella, ja siten sen arvo on väliltä 0 ja 1. Binaarimuunnelma on yleisin. Se määritellään kaavalla C / (A+B–C) missä C on yhteisten asetettujen bittien lukumäärä, A on sormenjälkeen A asetettujen bittien lukumäärä ja B on sormenjälkeen B asetettujen bittien lukumäärä. Useimpien sormenjälkien kohdalla yli 0,7 tai 0,8 vastaavuus osoittaa, että molekyyleillä on todennäköisesti samankaltaiset biologiset ominaisuudet (engl. *similar property principle*). Alle 0,3 kerroin ilmaisee, että molekyyleillä ei ole samankaltaisia ominaisuuksia.

Yksi aiheeseen liittyvä mitta on kosinikerroin, joka kertoo kahden vektorin välisen kulman. Ei-binäärisessä tapauksessa (ts. käyttämällä ei-binäärisiä deskriptoreja) Tanimoto on vektorien pistetulo jaettuna sormenjälkien A ja B summalla sekä miinustamalla pistetulo (C/(A+B–C)). Tästä syntyy Jaccard-indeksi binäärisissä sormenjäljissä.

Toinen yleinen mitta on euklidinen etäisyys, joka on teknisesti etäisyyden, ei samankaltaisuuden mitta. Euklidinen etäisyys on yksinkertaisesti Pythagoran etäisyys kahden pisteen välillä moniulotteisessa tilassa.

Tämä on erityisen hyödyllistä silloin, kun mittaamisen on noudatettava kolmioepäyhtälöä (ts. se on etäisyysfunktio), vaikka Soergel-etäisyyden (1-Tanimoto) on äskettäin osoitettu noudattavan kolmioepäyhtälöä positiivisissa deskriptoreissa. Huomaa, että binäärisissä sormenjäljissä euklidinen etäisyys on Hammingin etäisyyden neliöjuuri. Willet ym. (1998) ovat laatineet samankaltaisuuksien mittaamisesta erinomaisen yhteenvedon.

Tehtävät

10. Millaisia haasteita sanakirjaan pohjautuvat sormenjäljet asettavat samaan kemialliseen sarjaan kuuluvien samankaltaisten molekyylien vastaavuuksien laskennalle?
11. Missä olosuhteissa kahdella samankaltaisella eri molekyylillä esiintyy Tanimoto-vastaavuutta arvolla 1,0?
12. Voiko deskriptorityyppejä (esim. fragmentaarinen ja fysikaaliskemiallinen) yhdistää yksittäisessä sormenjäljessä? Jos kyllä, niin miten?
13. Mene PubChemin sivulle Atorvastatin.[20] Klikkaa linkkiä “Similar Compounds” kohdasta “Related Compounds”. Miksi hakutulosten kärjessä ei ole monia muita Statin-perheen lääkeaineita?[21]

[20] http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=60823

[21] http://en.wikipedia.org/wiki/Statin

4 2D-RAKENTEIDEN ETSIMINEN JA TALLENTAMINEN TIETOKANTAAN

Teeman tavoitteena on

- ymmärtää kemiallisen rakenneinformaation tallentamiseen liittyvät rajoitteet tiedostojen, laskentataukoiden ja tietokantojen tapauksessa
- oppia kemiallisten rakenteiden käsittelyyn soveltuvat tietokantatekniikat
- ymmärtää rakenteiden, alirakenteiden ja samankaltaisuuden hakeminen
- oppia hakemaan kemiallista informaatiota PostgreSQL- ja CHORD-tekniikoilla esimerkkien kautta.

Johdanto

Tähän mennessä on käsitelty 2D-rakenteiden esittäminen linjanotaatioiden, tiedostomuotojen ja sisäisten esitysmuotojen avulla. Tässä luvussa tietoa syvennetään tarkastelemalla, miten kemiallisten rakenteiden joukkoja tallennetaan, ja miten niitä voidaan hakea ja käyttää.

Askel tiedostotallentamisesta eteenpäin

Myös tiedostot mahdollistavat isojen informaatiomäärien tallentamisen, ja ne ovat paljon käytetty tallennusmuoto. Erityisesti lineaarinotaatiot mahdollistavat kemiallisten rakenteiden tallentamisen tiedostoihin, sillä ne koostuvat tekstistä. Esimerkiksi voidaan luoda SMILES-tiedosto, joka sisältää SMILES-kaavan lisäksi yhdisteen nimen ja jonkin ominaisuuden, kuten tietyn biologisen aktiivisuuden. Kentät voidaan erottaa toisistaan sarkaimella.

```
c1ccccc1        Benzene         3.6
c1cc(Cl)ccc1    Chlorobenzene   5.8
c1cc(Br)ccc1    Bromobenzene    2.4
```

Myös InChI-kaavalla voidaan luoda samanlaisia tiedostoja. Tiedostomuoto voi olla SD-, CML- tai moni muukin. Tiedoston täytyy vain sallia saman tiedon tallentaminen useaan kertaan, jotta voidaan tallentaa kemiallisten yhdisteiden "joukko".

Tiedon tallentamisen seuraava askel olisi tallentaa molekyylidata laskentataulukoihin tai jopa relaatiotietokantoihin. Yllä esitetyssä esimerkissä tietoa voitaisiin etsiä yhdisteen nimen perusteella, järjestää nimen ja ominaisuuden mukaan jne. Relaatiotietokannassa ei-rakenteelliselle tiedolle voitaisiin tehdä monimutkaisiakin kyselyjä. Haasteena on, että kemialliseen rakenteeseen pohjautuva hakeminen omaa useita rajoituksia. Esimerkiksi vain ensimmäisen seuraavista kyselyistä voidaan toteuttaa, ja sekin edellyttää sekä yhdistedatan että kyselyn kanonisointia.

1. Esiintyykö SMILES-kaavassa jokin tietty rakenne?

2. Etsi kaikki rakenteet, jotka sisältävät tiatsolidiinidionin (engl. *thiazolinedione*).
3. Etsi kaikki kyselyä vastaavat rakenteet.

Tämän tyyppisille kyselyille on käytettävä kemiaan erikoistuneita kyselyitä. Niiden käyttö edellyttää kemiaan erikoistuneita tietokantateknologioita.

Tietokantateknologiat

Vuosituhannen vaihteeseen saakka ainoastaan kemiaan erikoistuneet tietokannat mahdollistivat kemiallisen informaation hakemisen. Yksi ensimmäisistä oli MDL MACCS (1979), joka tarjosi 2D-rakenteiden tietokantojen tallentamista ja hakemista, mutta ei juurikaan muuta. Sen käytettävyys kasvoi vuonna 1985, kun se integroitiin ISIS-pakettiin. Integrointi salli kemiallisen rakenneinformaation tallentamiseen tietokantaan ja ei-rakenteellisten tietojen tallentamisen erilliseen Oracle-tietokantaan, sekä ISIS HOST -hakujärjestelmän näiden välille. Tietokoneille kehitettiin rajapintoja, jotka mahdollistivat rakenteiden piirtämisen ja visualisoimisen sekä kuvaajien pohjalta tehdyt tietokantakyselyt.

Järjestelmä oli tyypillisesti asiakas-palvelin-arkkitehtuurin mukainen, jossa palvelin vastaanottaa kyselyjä ja lähettää tuloksia asiakaskoneiden sovelluksille. Huomaa, että tämä edellyttää käyttötarkoitukseen erikoistuneiden ohjelmistojen asentamista sekä asiakaskoneisiin että palvelimelle. Pian kehitettiin myös monia muita samaan arkkitehtuuriin perustuvia hallintajärjestelmiä, kuten esimerkiksi Daylight Merlin / Thor ja Tripos Unity. Erikseen kehitettiin vain asiakaspuolella toimivia järjestelmiä (esim. Accord for Excel), joilla koneiden rakenneluetteloita käsiteltiin paikallisesti.

Suuri muutos alkoi Oraclen SQL-pohjaisen tietokannan hallintajärjestelmän version 8i julkaisusta, joka salli asiakkaalle omien sovelluskomponenttien kehittämisen. Omat sovellukset mahdollistivat tietokantojen toiminnallisuuksien laajentamisen asiakkaan tarpeiden mukaisesti. Pian ymmärrettiin, että 2D-rakenteiden tallentamiseen ja hakuun tarvittavat menetelmät voi-

daan toteuttaa laajennoksella, mikä mahdollisti yleisen Oracle SQL -tietokannan käyttämisen kemiallisten rakenteiden hakemiseen. Hyvä esimerkki Daylightin toteuttamasta laajennuksesta löytyy Jack Delaneyn MUG2000-puheesta.[22] Samankaltainen järjestelmä – Datablade – kehitettiin IBM:n Informix-tietokannalle, mutta sitä ei ole käytetty laajasti keminformatiikassa. gNovan CHORDiin on toteutettu samanlainen "plug-in" ilmaiseksi PostgreSQL-tietokannasta. Tällä hetkellä on saatavana monia Oracle- tai PostgreSQL-laajennuksia, kuten esimerkiksi:

- Accelrys Direct
- IDBS ActivityBase
- Daylight DayCart
- gNova CHORD
- ChemAxon JChem
- Open source OrChem ja ChemiSQL.

Huomaa, että kemian tietokantalaajennuksilla ei ole käyttöliittymää, joten sellainen tulee joko rakentaa tai paketoida tuote asiakasohjelmiston sisälle. Sekä Oracle and PostgreSQL ovat SQL-kieleen pohjautuvia relaatiotietokantoja, joiden käytöstä kemiassa O'Donnell (2009) on laatinut erinomaisen katsauksen.

Tiedeyhteisö on alkanut hyödyntää SQL:n lisäksi monia muitakin tietokantatekniikoita, mutta ne eivät sovellu toistaiseksi keminformatiikan tarpeisiin. Esimerkkejä vaihtoehdoista ovat Triplestore[23], NoSQL[24] ja JSON[25]. Erityisesti semanttinen Triplestore tarjoaa etuja suurten datamäärien integrointiin, linkittämiseen ja kartoittamiseen. Tätä on demonstroitu EU OpenPHACTS -projektissa.[26]

---

[22] Jack Delany – Daylight Chemistry Cartridge: http://www.daylight.com/meetings/mug00/Delany/cartridge.html

[23] http://en.wikipedia.org/wiki/Triplestore

[24] http://en.wikipedia.org/wiki/NoSQL

[25] http://en.wikipedia.org/wiki/JSON

[26] http://www.openphacts.org

Rakenne-, alirakenne- ja samankaltaisuushaku

Kemian 2D-rakenteiden tietokantahaussa käytetään yleisesti kolmea hakutyyppiä, joilla voidaan vastata kolmeen kysymykseen tai tiedontarpeeseen.

- Rakennehaku: Löytyykö tämä rakenne tietokannasta?
- Alirakennehaku: Etsi kaikki rakenteet, jotka sisältävät tämän alirakenteen.
- Samankaltaisuushaku: Etsi kaikki rakenteet, jotka ovat samanlaisia kuin tämä.

Yllä mainittuja hakuja voidaan yhdistellä tavallisiin numeroja tekstihakuihin, kuten esimerkiksi ”etsi kaikki rakenteet, jotka sisältävät tämän alirakenteen, ovat aktiivisia tällä määrittelyllä ja joiden logP < 5”.

Jos käytetään kanonisia SMILES-kaavoja ja kyselyitä, jotka on määritelty vastaavilla kanonisointialgoritmeilla, niin rakennehaku on yhtä yksinkertainen kuin SMILES-kaavan tekstihaku (vaikka pienimuotoiset variantit rajautuvat pois kanonisoinnin myötä). Luotettavampi alirakennehaku voidaan suorittaa graafisella isomorfismialgoritmilla. Alirakenteen haku vaatii subgraafisen isomorfismialgoritmin, kuten Ullman-algoritmin, käyttöä. Samankaltaisuushaku suoritetaan käyttämällä samankaltaisuus- tai etäisyyskertoimia ja deskriptoreita, kuten aiemmin on esitetty.

Rakenteiden ja samankaltaisuuksien etsimistä voidaan nopeuttaa esiseulomalla dataa fragmentaarisilla deskriptoreilla, johon ne alun perin suunniteltiinkin. Menetelmällä voidaan sulkea pois yhdisteitä, joilla ei ole kyselyssä haettuja ominaisuuksia (alirakennehaku), ja sellaisia, joiden suurin samankaltaisuus (perustuu olemassa oleviin ominaisuuksiin) kyselyyn olisi suurempi kuin määritelty raja (samankaltaisuushaku).

Alirakenteiden hakeminen SMARTS-kaavalla

Yksi kysymys on, että millaisilla kyselyillä voidaan hakea alirakenteita. Monien alirakenteiden kohdalla voidaan käyttää yksinkertaisesti SMILES-kaavaa aivan kuin alirakenne olisi täysrakenne. Usein alirakenteisiin halutaan kuitenkin lisätä ominaisuuksia, joita rakenteissa ei tavallisesti ole, kuten esimerkiksi kiinnityspisteiden määrittäminen, sidokset määrittelemättömiin atomeihin ja epäselvyys atomi- ja sidostyypeistä. Tiedontarpeena voi esimerkiksi olla sellaisen rengassysteemi löytäminen, joka kiinnittyy toiseen molekyyliin vain tietyistä kohdista.

Onneksi tämän tekemiselle on useita tapoja. Esimerkiksi MDL MOL/SD-tiedosto on laajennettavissa edustamaan kyselyn ominaisuuksia. Erityisen huomionarvoinen on SMARTS.[27] Se on SMILESin ylijoukko (engl. *superset*), joka on tarkoitettu kyselyjen laatimiseen. Yksinkertainen esimerkki SMARTSista olisi ***C(=O)O**, karboksyylihappo, joka eroaa SMILESista vain siinä, että tähti osoittaa kiinnityskohdan. Itse asiassa SMARTS sisältää laajan valikoiman lisämerkkejä kyselyiden esittämiseen. Hyödyllisiä Daylight-verkkosivuston resursseja ovat SMARTS-tutoriaali[28], SMARTS-esimerkit[29] ja SMARTS-harjoitukset[30] DepictMatch -työkalulla[31].

[27] http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

[28] http://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html

[29] http://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html

[30] http://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_practice.html

[31] http://www.daylight.com/daycgi_tutorials/depictmatch.cgi

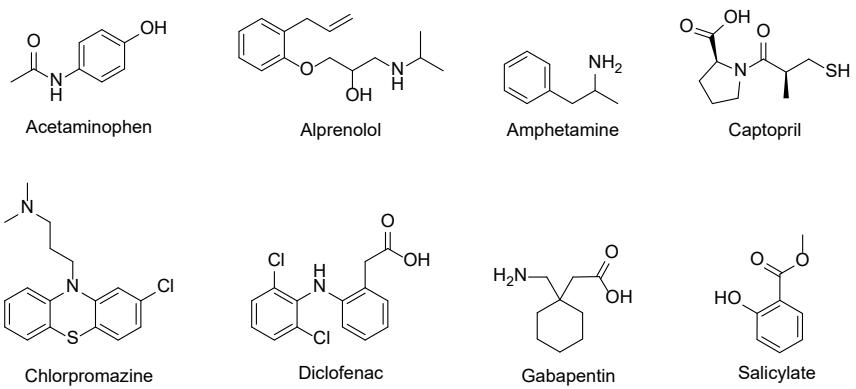
Asiakaspuolen tietokantarajapinnat

Asiakaspuolella tarvitaan jonkinlainen rajapinta tietokannoista etsimiseen. Tämä voi olla koneen rajapinta (esim. JDBC, ODBC, SOAP- tai REST-palvelu) tai ihmisen rajapinta (HTTP tai asiakaspuolen sovellus). Tietokantaan pääsyn lisääminen yhden ihmisen käyttöliittymän kautta on vanhentunut menetelmä. Palvelukeskeiset arkkitehtuurit antavat huomattavasti suuremman joustavuuden hakujen suorittamiselle erilaisissa sovelluksissa ja mashupeissa.

Asiakaspuolen rajapinnat tarvitsevat menetelmän 2D-rakenteiden visualisoimiseksi ja piirtämiseksi. Tämä voidaan toteuttaa useilla työkaluilla (esim. CDK tai OE-Chem) ja sovelluksilla ja lisäosilla (esim. Chemdraw Plugin tai JME). Se voidaan toteuttaa jopa REST-palvelulla.

Tietokantaesimerkki: PostgreSQL ja CHORD

Tässä esimerkissä työskennellään yleisistä lääkeaineista koostuvan pienen molekyyliaineiston kanssa.



Kuva 5. Esimerkissä työstettävät yhdisteet.

Seuraavissa esimerkeissä käytetään gNova CHORD -sovellusta PostgreSQL:n kanssa. Näiden esimerkkien ymmärtämiseksi tarvitset perustiedot SQL:stä.

Ensin luodaan uusi tietokanta SMILES, Name, LogP (oktanoli/vesi-jakautumiskerroin) ja sormenjälki jokaiselle kemialliselle rakenteelle:

```
create table gnovatest (smiles VARCHAR(200), name VARCHAR(50),
logp real, fkey BIT(166));
```

Seuraavaksi tietokantaan lisätään kenttäarvot 8 yhdisteelle:

```
INSERT INTO gnovatest (smiles, name, logp) VALUES (
'CC(=O)Nc1ccc(O)cc1', 'Acetaminophen', 0.27 );
INSERT INTO gnovatest (smiles, name, logp) VALUES (
'CC(C)NCC(O)COc1ccccc1CC=C', 'Alprenolol', 2.81 );
INSERT INTO gnovatest (smiles, name, logp) VALUES (
'CC(N)Cc1ccccc1', 'Amphetamine', 1.76 );
INSERT INTO gnovatest (smiles, name, logp) VALUES (
'CC(CS)C(=O)N1CCCC1C(=O)O', 'Captopril', 0.84 );
INSERT INTO gnovatest (smiles, name, logp) VALUES (
'CN(C)CCCN1c2ccccc2Sc3ccc(Cl)cc13', 'Chlorpromazine', 5.20 );
INSERT INTO gnovatest (smiles, name, logp) VALUES (
'OC(=O)Cc1ccccc1Nc2c(Cl)cccc2Cl', 'Diclofenac', 4.02 );
INSERT INTO gnovatest (smiles, name, logp) VALUES (
'NCC1(CC(=O)O)CCCCC1', 'Gabapentin', -1.37 );
INSERT INTO gnovatest (smiles, name, logp) VALUES (
'COC(=O)c1ccccc1O', 'Salicylate', 2.60 );
```

Seuraavaksi luodaan gNovapublic166keys-toiminnolla sormenjäljet SMILES-kentästä ja tallennetaan ne fkey-sormenjälkikenttään yhdistekohtaisesti:

```
update gnovatest set fkey = public166keys(smiles);
```

SQL select -komennolla saadaan kaikki tiedot näkyviin (vastaukset harmaalla):

```
select smiles,name,logp from gnovatest; select * from gnovatest;

          smiles             |      name      | logp |
fkey
-----------------------------+----------------+------+-----------------------------------
-----------------------------------------------------------------------------------------
---------------------------
 CC(=O)Nc1ccc(O)cc1              | Acetaminophen  |  0.27 |
00000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000001000000000000000000100100010
00000000000010101000100010000000110101111111111110
 CC(C)NCC(O)COc1ccccc1CC=C       | Alprenolol     | 2.81 |
00000000000000000000000000000000001000000000000000000100000
00000000000010100000001000000010000101011000100001000100100
00000001010011000000100011000011111011111111111110
 CC(N)Cc1ccccc1                  | Amphetamine    | 1.76 |
00000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000010000000000000000000100000010001000
00000000000000000000000000000000000110101111010
 CC(CS)C(=O)N1CCCC1C(=O)O        | Captopril      | 0.84 |
00000000000000000000000000000000000000000000000000000000000000000000
00000000000000010000011010010111001100010000001011000101O
00111000010010000111100000011101001111111100110
 CN(C)CCCN1c2ccccc2Sc3ccc(Cl)cc13 | Chlorpromazine | 5.2 |
00000000000000000000000000000000000010000000000000000000001
00000000000000110000110001011000010000101001010110010001100
01110000000000011010000101101110001010110101111010
 OC(=O)Cc1ccccc1Nc2c(Cl)cccc2Cl  | Diclofenac     | 4.02 |
00000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000001001000000000000011001000000000000
00001000000011111100010000110000110011111110111110
 NCC1(CC(=O)O)CCCCC1             | Gabapentin     | -1.37 |
00000000000000000000000000000000000000000000000000000000000000000000
00000010000000000000001010000011000000001000100000001000000
00001000011011000000100000001000001110111010111O
 COC(=O)c1ccccc1O                | Salicylate     |  2.6 |
00000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000001000100000000000000000100000
00001001000000000000100011010001010100101101111O
```

Nyt kun tietokanta on täynnä, siitä voidaan tehdä hakuja SQL-kyselyillä ja keminformatiikkaan erikoistuneilla toiminnoilla. Esimerkiksi voidaan hakea tietyn nimistä yhdistettä tai LogP-arvoa tietyistä rajoista. Seuraavalla tavalla tehtäisiin alirakenteiden haku karboksyylihapporyhmän sisältäville rakenteille (huomaa karboksyylihapon SMARTS-esitys):

```
select smiles,name,logp from gnovatest where matches(smiles,
'*C(=O)O');

            smiles             |    name    | logp
-------------------------------+------------+------
 CC(CS)C(=O)N1CCCC1C(=O)O      | Captopril  |  0.84
 OC(=O)Cc1ccccc1Nc2c(Cl)cccc2Cl | Diclofenac |  4.02
 NCC1(CC(=O)O)CCCCC1           | Gabapentin | -1.37
 COC(=O)c1ccccc1O              | Salicylate |  2.6
(4 rows)
```

Tarkennetaan hakua palauttamalla vain karboksyylihapporyhmän sisältämät yhdisteet, joiden LogP > 1:

```
select smiles,name,logp from gnovatest where (matches(smiles,
'*C(=O)O') AND (logp>1.0));

            smiles             |    name    | logp
-------------------------------+------------+------
 OC(=O)Cc1ccccc1Nc2c(Cl)cccc2Cl | Diclofenac | 4.02
 COC(=O)c1ccccc1O              | Salicylate |  2.6
(2 rows)
```

Seuraavassa kyselyssä suoritetaan SMILES-kaavoilla esitetty samankaltaisuushaku aspiriinille. Kysely palauttaa vain ne yhdisteet, joiden sormenjälkien Tanimoto-samankaltaisuus on suurempi kuin 0,6. Huomaa, että tässä tapauksessa palautui vain yksi yhdiste – yllättäen salisylaatti, joka on aspiriinin prekursori:

```
select smiles,name,logp from gnovatest where tanimoto(fkey, pub-
lic166keys('CC(=O)Oc1ccccc1C(=O)O')) > 0.6;

     smiles      |    name    | logp
-----------------+------------+------
 COC(=O)c1ccccc1O | Salicylate | 2.6
(1 row)
```

Kun työskentely tietokannan kanssa on valmis, voidaan se poistaa:

```
drop table gnovatest;
```

Vapaat ja avoimet kemian tietokannat

Verkossa on useita julkisia kemiallisia rakennetietokantoja, joista merkittävimpiä ovat PubChem, ChemSpider, ChEMBL ja eMolecules.

PubChem – http://pubchem.ncbi.nlm.nih.gov – on tietokanta, joka sisältää informaatiota kymmenistä miljoonista yhdisteistä, mukaan lukien kasvavan määrän bioaktiivisuustietoja. Yksinkertaiset haut voidaan suorittaa avainsanoilla, ja edistyneemmät haut kemiallisen rakenteen (rakenne-, alirakenne- ja samankaltaisuushaku) ja muiden tekijöiden perusteella.

Hakutulosten klusterointiin ja bioaktiivisuuden analysointiin on tarjolla kehittyneitä työkaluja. Niiden käytön tueksi löytyy monenlaisia koulutusmateriaaleja, kuten esimerkiksi PubChem-ohjeet[32], NCBI:n järjestämien edellisten PubChem-kurssien materiaalit[33] ja Berkeleyn yliopiston (Kalifornia, USA) koulutusvideoita[34].

[32] http://pubchem.ncbi.nlm.nih.gov/help.html

[33] https://www.ncbi.nlm.nih.gov/Class/PubChem/course.html

[34] http://www.lib.berkeley.edu/CHEM/instruction/pubchem

Chemspider – http://www.chemspider.com – on ilmainen kemian tietokanta. Se on vaihtoehto PubChemille sisältäen myös spektroskopiadataa. Sivusto tarjoaa useita koulutusmateriaaleja käytön tueksi.

ChEMBL – https://www.ebi.ac.uk/chembl – on bioaktiivisten lääkkeiden kaltaisten pienten molekyylien tietokanta. ChEMBL sisältää 2D-rakenteita, laskettuja ominaisuuksia (esim. LogP, molekyylipaino, Lipinski-parametrit jne.) ja valittuja tietoja bioaktiivisuudesta (esim. sitoutumisvakiot, farmakologia ja ADMET-tiedot).

eMolecules – http://www.emolecules.com – on tietokanta, joka keskittyy kaupallisesti saatavilla oleviin yhdisteisiin.

Tehtävät

14. Mikä on ainoa erikoistunut keminformatiikan haku, joka on mahdollista tavallisella laskentataulukolla tai tiedostolla?
15. Kuvaile omin sanoin, mitä oheinen SMARTS tarkoittaa ja millaista fragmenttia se edustaa: **[#6][CX3](=O)[#6]**.
16. Selvitä, kuinka monta erillistä kemiallista rakennetta on saatavana PubChem- ja ChemSpider-tietokannoissa.

5 KEMIALLISTEN REAKTIOIDEN KÄSITTELY TIETOKONEELLA

Teeman tavoitteena on

- ymmärtää erityyppisten kemiallisten reaktioiden yhteys rakenteiden esittämiseen
- oppia tuntemaan reaktio-tietokantojen ja synteesien suunnittelujärjestelmien väliset erot
- oppia käyttämään joitain reaktiotietokantojen edistyneitä hakutoimintoja.

Kemialliset reaktiot

Kemialliset reaktiot ovat prosesseja, joissa aineet muuttuvat toisiksi aineiksi. Reaktiot on jaoteltu reaktiotyyppeihin, kuten esimerkiksi korvautumisreaktiot tai happoemäsreaktiot. Lääkeaineiden kehittämisessä orgaaniset reaktiot ovat erityisen kiinnostavia. Reaktioihin voidaan liittää monenlaista informaatiota, mukaan lukien

- reaktioyhtälö ja stoikiometria
- yksityiskohtainen reaktiomekanismi
- reaktioon liittyvät katalyytit ja liuottimet
- reaktio-olosuhteet (sisältää usein numeerista ja tekstiinformaatiota)
- reaktiosaanto (numeerinen).

Huomaa, että reaktio voi olla geneerinen (ts. sovellettavissa moniin yhdisteisiin, ei vain yhteen tiettyyn) tai spesifinen, ja reaktioyhtälö saattaa esittää vain yksityiskohtaisen reaktiomekanismin yksinkertaistetut alku- ja loppupisteet. Yksinkertaisista orgaanisista reaktioista löytyy tietoa Wikipediasta[35] tai mistä tahansa yliopistotason orgaanisen kemian oppikirjasta.

Reaktioyhtälöt esitetään yleensä paperilla kuten matemaattiset yhtälöt, joissa reagenssit ovat vasemmalla ja tuotteet oikealla. Tavanomaisesti +-merkkiä käytetään erottamaan saman puolen aineet toisistaan. Nuoli erottaa lähtöaineet tuotteista ja osoittaa reaktion suunnan.

Keminformaattisesta näkökulmasta tärkein huolenaihe on yksittäisten kemiallisten rakenteiden esittäminen sekä lähtöaineiden ja tuotteiden välisen suhteen kuvaaminen (tai reaktiomekanismin, jos se tallennetaan). Kaikki muut tiedot esitetään triviaalisti (esim. teksti, numeerinen jne.). Huomaa, että muutos on reaktiomekanismin sisältävän informaation osajoukko, ja on mahdollista esittää muutoksia, jotka eivät vastaa valideja reaktiomekanismeja.

[35] https://fi.wikipedia.org/wiki/Luokka:Orgaaniset_reaktiot

Joka tapauksessa, informaation esittäminen tietokoneella edellyttää usein yksityiskohtaisempaa tietoa muutoksesta kuin sen esittäminen paperilla (esim. eksplisiittiset rakenteet, yhden rakenteen kartoitus toiseen). Tietokantaan voidaan tallentaa muutos tai reaktiomekanismi tai ne molemmat.

Tähän mennessä on opittu, kuinka 2D-rakenteita voidaan esittää SMILES-, InChI-, MDL MOL-tiedostoilla jne. Lisäksi tarvitaan menetelmä, joka kartoittaa reaktion lähtöaineista tuotteisiin ja esittää siten koko reaktion. Tämä voidaan toteuttaa reaktio-SMILES- ja SMIRKS-kaavoilla (ks. luku 2).

Reaktio-SMILES laajentaa SMILES-kaavaa siten, että se tukee lähtöaineiden ja tuotteiden tunnistamista sekä niiden välisen muutoksen kartoitusta. SMIRKS on vielä edistyneempi sallien myös sellaisten muutosten määrittelyn, jotka eivät koostu täydellisistä rakenteista vaan SMARTSin kaltaisella tavalla kuvatuista fragmenteista. Liitostaulukkoon voidaan lisätä erilaisia reaktiotietoja – esimerkiksi MDL MOL/SD-tiedoston laajennus RXN-tiedosto määrittelee rakenteet reagensseiksi tai tuotteiksi. SD-tiedosto voi sisältää useita MOL-tiedostoja, ja samalla tavalla RD-tiedosto voi sisältää useita RXN-tiedostoja.

Huomaa, että InChI:lle ei tällä hetkellä ole tarjolla reaktioSMILES- tai SMIRKS-tiedostoja vastaavia esitysmuotoja.

Reaktiotietokannat

Historiassa on julkaistu useita kirjoja, jotka indeksoivat ja kuvaavat orgaanisia reaktioita. Näistä tunnetuimpia on teos nimeltä Beilstein Handbook of Organic Chemistry[36]. Teos siirtyi hitaasti Beilstein-tietokantaan, joka on nyt yksi suurimmista reaktioarkistoista, jota Elsevier jakaa osana Reaxsys-järjestelmäänsä[37]. Muita tietokantoja ovat Chemical Abstract Service CASREACT[38]

[36] https://en.wikipedia.org/wiki/Beilstein_database

[37] https://www.reaxys.com

[38] http://www.cas.org/expertise/cascontent/casreact.html

ja SPRESI[39]. Lisäksi on olemassa joitain ilmaisia tietokantoja kuten The Chemical Thesaurus[40] ja WebReactions[41].

Reaktiotietokannoista on tärkeää tietää kaksi asiaa. Ensinnäkin ne rakentuvat reaktiodatan lähteiden ympärille. Yleisin datalähde on tieteellinen kirjallisuus, josta reaktiot on poimittu manuaalisesti. Toiseksi ne eroavat synteesisuunnittelujärjestelmistä (*CAESA*[42] ja *WODCA*[43]), jotka tukevat kemistejä reaktioiden suunnittelussa esimerkiksi tiettyjen sääntöjen avulla, mutta eivät välttämättä sisällä yksityiskohtaista reaktiotietokantaa.

Moniin yksinkertaisiin rakennetietokantoihin on kehitetty räätälöityjä järjestelmiä ja käyttöliittymiä, joihin sisältyy usein sekä yhdisteiden että reaktioiden etsiminen (esim. DiscoveryGate ja SciFinder). Nämä järjestelmät on kehitetty erityisesti kemiaan erikoistuneiden informaatikkojen ja synteesikemistien käyttöön. Järjestelmien käyttöä opetetaan kemian tiedonhankinnan kursseilla.

Reaktiotietokanta voidaan toteuttaa tallentamalla SMILES-kaava tietokannan tekstikenttään, jonne voidaan tallentaa myös reaktioSMILES- tai SMIRKS-kaava. Suurin osa kemian tietokantasovelluksista osaa toimia niiden kanssa. Seuraavilla sovelluskomponenteilla voidaan käsitellä reaktioita:

- MDL Isentris
- Tripos Auspyx
- Daylight DayCart
- Accelrys Accord
- IDBS ActivityBase
- ChemAxon JChem
- gNova CHORD.

[39] http://www.spresi.com

[40] http://www.chemthes.com

[41] http://www.openmolecules.org/webreactions/index.html

[42] http://www.keymodule.co.uk/products/caesa/index.html

[43] http://www2.chemie.uni-erlangen.de/software/wodca

Suorien rakenne-, alirakenne- ja samankaltaisuushakujen lisäksi reaktiotietokantojen tulee tukea myös erikoistuneempia hakuja. Usein haku halutaan rajata vain joko lähtöaineisiin tai tuotteisiin. Halutaan esimerkiksi löytää kaikki tuotteisiin liittyvät reaktiot. Tämä tarve johtaa nopeasti edistyneempiin hakuihin, kuten esimerkiksi:

- kaikki reaktiot, jotka sisältävät tietyn alirakenteen
- eri synteesireitit tietylle nimirektiolle (esim. etsi kaikki Dies–Alder-reaktiot)
- kaikki reaktiot, jotka vastaavat haun reaktiota
- reaktioketjujen löytäminen lähtöaineiden joukosta, joita voidaan käyttää tietyn rakenteen syntetisoimiseen.

Kyselyissä tulee voida käyttää sekä tekstiä että numeerisia hakuja, joten reaktiotietokantahaut voivat olla melko monimutkaisia. Lisäksi reaktioketjujen löytäminen on erittäin haastavaa, koska ne on esitetty kuvilla.

Tehtävät

17. Selitä, mitä eroa on reaktioyhtälöllä, reaktiomekanismilla ja kemiallisella muutoksella.
18. Käy läpi WebReactions-tutoriaali[44]. Tutoriaalissa määritettiin kysely. Oliko se reaktioyhtälö, reaktiomekanismi vai kuvasiko se kemiallista muutosta?
19. Luo SMIRKS-jono kyselylle, jonka määritit WebReactions-tutoriaalissa.
20. Miksi luulet, että on olemassa vain muutamia ilmaisia ja avoimia reaktiotietokantoja?

[44] http://www.openmolecules.org/webreactions/tutorial.html

6 3D-RAKENTEIDEN ESITTÄMINEN TIETOKONEELLA

Teeman tavoitteena on

- tutustua 3D-rakenteiden informaatiolähteisiin
- ymmärtää konformaatiojoustavuuden kuvaamiseen liittyvät haasteet ja kaksi siihen kehitettyä ratkaisua
- hallita koordinaatiotaulukot ja etäisyysmatriisit
- oppia, mikä on 3D-farmakofori ja kuinka sitä voidaan käyttää tietokantahaussa.

3D-rakennedatan tuottaminen

2D-rakenteet voidaan generoida atomien ja niiden sitoutumisen avulla. Nämä yleissäännöt pätevät kaikkiin yhdisteisiin, joten 2D-rakenteiden alkuperää ei tarvitse tarkemmin pohtia. Mutta yhdisteen 3D-rakenteesta ei ole olemassa a priori -tietoa, joka määräisi 3D-rakenteen. Lisäksi kaikki yhdisteet ovat jossain määrin joustavia, joten niiden 3D-rakenne muuttuu jatkuvasti. On myös pidettävä mielessä, että sekä 3D- että 2D-rakenteiden kohdalla on kyse mallista eikä todellisuudesta. Mallit pohjautuvat tähän mennessä parhaaseen tietoon suuren mittakaavan sumeasta kvantti-ilmiöstä.

3D-rakenneinformaatioon on kolme päälähdettä: kaksi kokeellista menetelmää (röntgenkristallografia ja NMR-spektroskopia) sekä yksi laskennallinen (tietokoneella luodut 3D-rakenteet). Kokeellisia menetelmiä ei tässä teoksessa käsitellä tarkemmin. Todetaan vain, että molemmat tuottavat yhdisteelle tietyssä muodossa joukon atomikoordinaatteja (esimerkiksi kidemuodon röntgenrakenteille). Koordinaatit eivät välttämättä vastaa muotoa, johon yhdiste taipuu esimerkiksi sitoutuessaan proteiinikohteeseen. Tämän vuoksi tarvitaan menetelmiä molekyylin taipuisuuden käsittelyyn.

Konformaatiojoustavuus

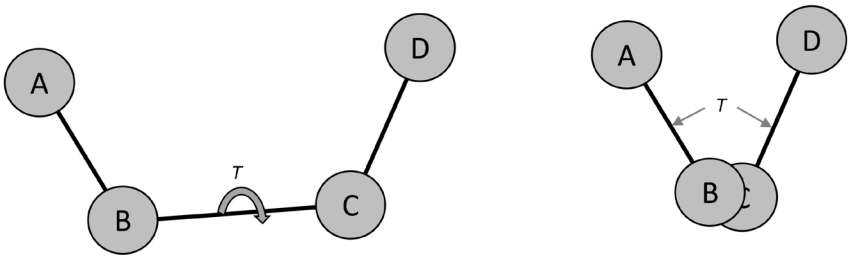
Useimmissa yhdisteissä on kiertyviä sidoksia, joiden vuoksi molekyyli voi taipua moniin erilaisiin konformereeihin. Tämän vuoksi molekyylillä ei ole vain yhtä tiettyä 3D-rakennetta, vaan jokaisella on ääretön määrä (tai vähemmän kuin ääretön, jos kiertoyksiköiden oletetaan olevan diskreettejä) mahdollisia konformeereja. Kaikkien konformeerien energiat eivät ole yhtä suuria, ja molekyylit taipuvat mieluimmin alhaiseen kuin korkeaan energiatilaan.

Konformaatiojoustavuus voidaan ottaa huomioon joko tallentamalla vain yksi alhaisen energian omaava konformeeri ja antamalla algoritmien taivutella molekyyliä tarpeen mukaan tai tuottamalla useita lähtökonformeereja, eli otokseen sisällytettäisiin useita alhaisen energian muotoja. Ratkaisusta riippuen konformaa-

tiojoustavuutta voidaan siis käsitellä joko esitystasolla tai algoritmitasolla. Esitystasolla tallennetaan useita konformeereja. Algoritmitasolla käytetään vain yhtä 3D-rakennetta, ja ohjelmistot huomioivat konformaatiojoustavuuden laskennallisesti ottamalla näytteet konformaatioavaruudesta.

Ennen konformaatiojoustavuuden käsittelyä on päätettävä, kuinka kiertyvät sidokset määritetään. Yksi toimiva nyrkkisääntö on, että mikä tahansa yksittäinen sidos voi kiertyä, joka ei ole osa rengasta, ei ole terminaalinen (esim. metyyli) ja ei ole osa konjugoitua systeemiä (esim. amidi). Sääntö ei kuitenkaan ole täydellinen. Tiedetään, että konjugoidun systeemin sidokset voivat pyöriä tietyn asteen verran (konjugaatioaste), ja renkaat voivat taipua (esim. sykloheksaanin tuoli- ja venemuotojen kiertyminen).

Sidosten kiertymisen yhteydessä englanninkielisessä kirjallisuudessa käytetään termejä *torsion angle* ja *dihedral angle*.[45] Termit ovat synonyymejä ja viittaavat A–B-sidosten ja C–D-sidosten väliseen suhteelliseen asemaan tai kulmaan, kun tarkastellaan neljää peräkkäistä atomia A-B-C-D (kuva 6).



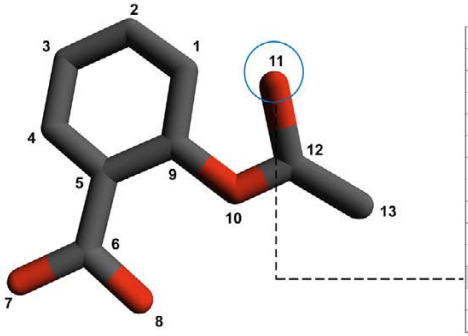
Kuva 6. Torsiokulma A–B-C–D.

[45] Suomentajan kommentti: Suomen kielen termi on torsiokulma.

3D-konformeerien esittäminen tietokoneella

2D-rakenne sisältää tiedon atomeista ja niiden sitoutumisesta, mutta 3D-rakenteeseen tarvitaan lisäksi atomien koordinaatit origoon suhteutettuna. Tämän tiedon tallentamiseksi ei ole vakiintunutta lineaarinotaatiota, vaikkakin Sybyl Line Notation (SLN) sallii atomien merkitsemisen koordinaateilla. Yleisempi tapa on liitostaulukon kaltainen tiedostomuoto, usein joko MDL MOL-, SD-tiedosto tai Sybyl MOL2-tiedosto. Myös muita tiedostomuotoja voidaan käyttää, kuten CML, PDB ja pelkästään koordinaateille tarkoitettu XYZ-tiedosto.

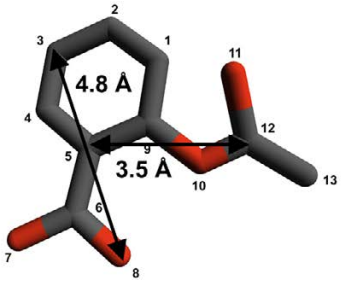
Yksi mahdollisuus on luoda koordinaattitaulukko, joka on yksinkertainen atomihakutaulukon laajennus. Laajennus tallentaa X-, Y- ja Z-koordinaatit jokaiselle atomille suhteessa origoon. Tämä koordinaattijärjestelmä perustuu usein ångströmeihin eli yksi yksikkö on yksi ångström (ks. kuva 7).



Atomi	Merkki	X	Y	Z
1	C	−1.8920	−0.9920	−1.5760
2	C	−1.3680	−2.1480	−0.9880
3	C	−0.0760	−2.1440	−0.4640
4	C	0.708	−0.9840	−0.5200
5	C	0.200	−0.1560	−1.1960
6	C	−1.1080	0.1600	−1.6520
7	O	2.0840	−1.0280	0.1040
8	O	2.5320	−2.0320	0.6360
9	C	2.8760	0.0240	0.1120
10	O	0.7520	1.3320	−1.0840
11	O	0.6680	2.0240	0.0320
12	C	1.3000	3.0600	0.1520
13	C	−0.2400	1.5760	1.4440

Kuva 7. Esimerkki koordinaattitaulukosta.

Koordinaattitaulukosta voidaan johtaa etäisyysmatriisi (ks. kuva 8), joka ilmaisee kahden atomin välisen etäisyyden ångströmeissä.



	1	2	3	4	5	6	7	8	9	10	11	12	13
1		1.4	2.4	2.8	2.4	3.8	4.8	4.2	1.4	2.4	2.7	2.9	4.3
2			1.4	2.4	2.8	4.3	5.1	5.0	2.4	3.7	3.9	4.2	5.6
3				1.4	2.4	3.8	4.2	4.8	2.8	4.2	4.7	4.9	6.4
4					1.4	2.5	2.8	3.6	2.4	3.7	4.7	4.6	6.1
5						1.5	2.4	2.3	1.4	2.3	3.7	3.5	4.8
6							1.3	1.2	2.5	2.8	4.4	3.9	5.0
7								2.2	3.7	4.1	5.7	5.2	6.3
8									2.8	2.5	4.2	3.5	4.3
9										1.4	2.6	2.3	3.7
10											2.2	1.3	2.5
11												1.2	2.4
12													1.5
13													

Kuva 8. Esimerkki etäisyysmatriisista.

Huomaa, että muodostuu täydellinen verkko.[46] Koordinaattitaulukoiden ja etäisyysmatriisien lisäksi yhdisteen 3D-konformeerien joustavuusaste voidaan spesifioida muillakin tavoilla. Esimerkiksi voidaan käyttää kahta koordinaattitaululukkoa, joista toiseen tallennetaan atomin X-, Y- ja Z-koordinaattien minimiarvot ja toiseen maksimit. Vastaavasti voidaan käyttää kahta etäisyysmatriisia.

3D-rakenteiden tuottaminen ja muokkaaminen tietokoneella

On olemassa useita ohjelmistoja, jotka konvertoivat 2D-rakenteet (esim. SMILES-kaavasta) 3D-rakenteiksi. Konvertoidut rakenteet ovat usein valideja, mutta eivät energiaminimoituja, ellei prosessiin ole yhdistetty energian minimointia. Ohjelmistot tuottavat yhden rakenteen (output-tiedosto) tai 3D-rakenteiden kokonaisuuden. Konvertointi toteutetaan usein sääntöihin pohjautuvina fragmentteina. 2D-rakennne jaetaan pieniksi fragmenteiksi, jotka sovitetaan ennalta määriteltyyn 3D-fragmenttien sanakirjaan. Fragmenteista luodaan sanakirjan sääntöjen avulla valmis 3D-rakenne teoriaan pohjautuen.

[46] Engl. *"Fully connected graph"* tai *"Complete graph"*

Konvertointimenetelmät voivat perustua myös etäisyysgeometriaan, jossa algoritmi poimii konformaatioavaruudesta valideja konformeerejä etäisyysrajojen perusteella. SMI23D on esimerkki etäisyysgeometriamenetelmästä. Se on Indianan yliopiston tuottama avoin ohjelmisto, joka konvertoi SMILES-merkkijonosta 3D-rakenteen SD-tiedostona. Voit kokeilla sitä osoitteessa http://cheminfov.informatics.indiana.edu/rest/thread/d3.py/SMILES/, jonka loppuun liitetään konvertoitava SMILES-merkkijono. Sovellus tarjoaa SD-tiedostoa automaattisesti ladattavaksi.[47]

```
http://cheminfov.informatics.indiana.edu/rest/thread/d3.py/SMILES/c1ccccc1
```

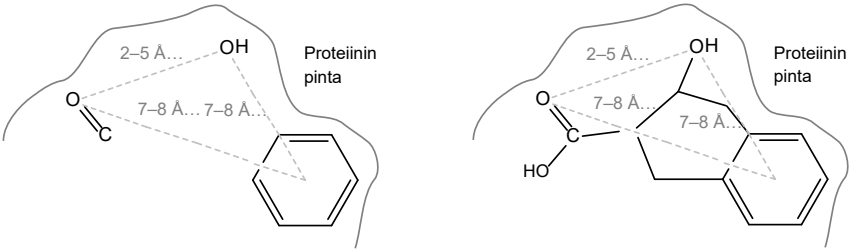
Useimmat 3D-rakenteiden generointimenetelmät suorittavat myös energian minimoinnin, jota voidaan soveltaa mistä tahansa lähteestä kerättyyn 3D-dataan (esim. röntgenkristallografia- tai NMR-data). Energian minimointialgoritmi ottaa konformeerin geometrian lähtöpisteeksi, jota se pyörittää ja taivuttaa minimoiden samalla rakenteen potentiaalienergian. Tämä voidaan tehdä millä tahansa optimointialgoritmilla. Jotkut algoritmit etsivät paikallisen minimin, kun toiset pyrkivät löytämään globaalin minimin.

3D-farmakoforit

Farmakofori on molekyylin niiden ominaisuuksien kuvaus, joita sen sitoutuminen tiettyyn proteiiniin edellyttää. Farmakoforilla kuvataan yleensä 3D-rakenteellisia ominaisuuksia, kuten esimerkiksi vety-sitoutumispotentiaalia. Teknisesti farmakofori voidaan määritellä ominaisuusjoukoksi ja ominaisuuksien etäisyysrajoiksi

[47] Suomentajan kommentti: Alkuperäisteoksessa oli linkki ohjelmiston lähdekoodiin, mutta se ei toiminut 10.7.2020. Lisätietoa ohjelmistosta löytyy sivulta: https://depth-first.com/articles/2007/12/12/simple-3d-conformer-generation-with-smi23d.

3D-avaruudessa. Farmakofori voidaan generoida joko sitoutumiskohteesta tai ligandeista. Esimerkiksi "OH-ryhmä, joka on 2–5 ångströmin päässä karboksyyliryhmän happiatomista, ja molemmat ovat 7–8 ångströmin päässä bentseenirenkaasta (kuva 9)":



Kuva 9. Esimerkkifarmakoforin visualisointi.

Farmakoforia voidaan käyttää myös tietokantakyselynä. Huomaa, että farmakoforihaku on alirakennehaun kaltainen siinä mielessä, että se on täydellisen verkon etäisyysmatriisin alikysely. Farmakofori voidaan esittää monella tavalla, kuten esimerkiksi farmakoforipisteiden etäisyysmatriisina sisältäen sanakirjan pistetyypeille, jotka voivat sisältää 3D-alirakenteiden koordinaatit tai 2D-ominaisuuksien SMARTS-merkkijonot. Huomaa, että usein on tarve esittää etäisyysväli, mutta ei tarkkaa etäisyyttä. Lisäksi farmakoforipisteet voivat olla moniselitteisiä, jotka on niin ikään kyettävä esittää.

3D-deskriptorit ja sormenjäljet

Aivan kuten 2D-rakenteiden kanssa, myös 3D-rakenteille ja ominaisuuksille voidaan luoda deskriptoreita. 3D-farmakoforit ovat 2D-rakenneavaimia vastaavia fragmentteja. Toisinaan niitä kutsutaan tripleteiksi tai kvarteteiksi fragmenttien sisältämien atomien lukumäärien perusteella. Huomaa, että nämä fragmentit voivat sisältää etäisyyksiä ja moniselitteisiä pisteitä aivan kuten farmakoforit. Molekyylijoukolle voidaan generoida valtava määrä triplet-

tejä ja kvartetteja, joten yleensä ne hajautetaan kiinteään bittilukumäärään.

3D-rakenteille voidaan luoda muun muassa atomipohjaisia deskriptoreita (esim. semiempiirisistä menetelmistä luodut osittaisvaraukset) tai molekyylipohjaisia (esim. sähköstaattiset, steeriset ja hydrofobiset kentät). Näitä voidaan hyödyntää monissa sovelluksissa, kuten esimerkiksi molekyylisuuntauksessa, telakoinnissa ja samankaltaisuusanalyyseissä.

3D-tietokantojen käyttö

NetSci[48] tarjoaa hyvän yleiskuvan siitä, miten 3D-rakennetietokantoja käytetään lääkeainetutkimuksessa.[49] Farmakoforinen haku toimii samoin kuin 2D-rakenteiden alirakennehaku, eli suoritetaan farmakoforihaku, joka palauttaa kaikki kyselyä vastaavat molekyylit. Teknisesti haku toteutetaan joko taivuttamalla molekyyliä tai varastoimalla useita konformeerejä.

3D-rakenteiden samankaltaisuushaku voidaan toteuttaa yksinkertaisesti laskemalla Tanimoto-kerroin tai kahden sormenjäljen välinen euklidinen etäisyys (kuten 2D-rakenteiden kanssa). Mutta 3D-samankaltaisuuden laskemiseen on olemassa myös muita tapoja, jotka eivät perustu 3D-samankaltaisuuteen. Esimerkiksi kahta molekyyliä voidaan verrata toisiinsa etäisyysmatriisilla tai asettamalla molekyylit päällekkäin ja laskemalla kenttien päällekkäisyysaste.

[48] https://web.archive.org/web/20110101081356/http://www.netsci.org/Science/Cheminform/feature06.html

[49] Suomentajan kommentti: NetSci-sivustoa ei enää ole. Viite 47 ohjaa verkkosivustojen arkistointipalveluun.

Esimerkkejä 3D-rakennetietokannoista

Kattavin röntgenkristallografisesti tuotettu tietokanta on Cambridge Structural Database.[50] Tammikuussa 2009 se sisälsi 469 611 rakennetta.[51] Tietokanta sisältää useita työkaluja rakenteiden tarkasteluun ja analysointiin, mukaan lukien useita ilmaisia ominaisuuksia.[52] Tarjolla on muun muassa ilmainen 500 yhdisteen osajoukko opetustarkoituksiin.

Myös PubChem sisältää 3D-rakenteita ja sallii monenlaisia hakuja. Palvelun 3D-ominaisuuksia on käsitelty kattavasti Journal of Cheminformatics -lehden PubChem3D-teemakategoriassa.[53]

Tehtävät

21. Osiossa esiteltiin kaksi tapaa käsitellä konformaatiojoustavuutta. Vertaile, millaisia perusteluja molemmille tavoille esitettiin.
22. Luo SMI23D-sovelluksella 3D-rakenteet klooribentseenille ja bromibentseenille. Mitkä ovat rakenteiden MMFF94-voimakentällä minimoidut energiat?
23. Missä tapauksessa farmakoforihaku voisi olla hyödyllisempi kuin 2D-alirakennehaku, jos tiedonhankinnan kohteena olisi lääkeaineiden kehittäminen?

[50] https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd

[51] Suomentajan kommentti: Vuonna 2020 rakenteita oli yli 1 000 000.

[52] https://www.ccdc.cam.ac.uk/Community/csd-community

[53] https://www.biomedcentral.com/collections/pubchem3d

7 KEMIALLISTEN RAKENTEIDEN ESITTÄMINEN VERKOSSA JA TIETEELLISISSÄ JULKAISUISSA

Teeman tavoitteena on oppia

- kemialliselle informaatiolle asetetut neljä vaatimusta, jotka tekevät siitä hyödyllistä verkossa ja asiakirjoissa
- asiakirjoissa olevien kemiallisten rakenteiden konelukemisen haasteet.
- InChI-kaavan mahdollisuudet tiedon löydettävyyden tukemisessa
- kontekstualisoinnin tärkeys.

Kemiallinen informaatio asiakirjoissa

Internetiä käytetään yhä enemmän kaikenlaiseen tiedonhankintaan. Samalla myös verkosta löytyvän tieteellisen tiedon määrä kasvaa koko ajan. Verkossa tieteellistä tietoa löytyy esimerkiksi elektronisista asiakirjoista ja lehtien arkistoista. Näistä lähteistä kemiallisen tiedon etsiminen on kuitenkin haastavaa. Jotta asiakirjojen sisältämä kemiallinen rakenneinformaatio olisi hyödyllistä, on sen oltava:

- **Koneluettavaa** – rakenteiden tulee olla sellaisessa muodossa, tietokone voi lukea ja ymmärtää niitä.
- **Löydettävää** – asiakirjoista tulee voida hakea kemiallisia yhdisteitä kyselyillä (esim. rakenne-, alirakenne- ja samankaltaisuushaut).
- **Saavutettavaa** – kun kyselyä vastaava asiakirja on löytynyt, on asiakirjan koko sisältöön päästävä tietokoneella esimerkiksi koneellista jatkokäsittelyä tai lukemista varten.
- **Kontekstualisoitua** – asiakirjan kemiallisella informaatiolla tulisi olla yhteys muuhun aiheeseen liittyvään relevanttiin informaatioon, kuten esimerkiksi biologisiin aktiivisuuksiin, kemiallisiin ominaisuuksiin, reaktiokuvauksiin jne.

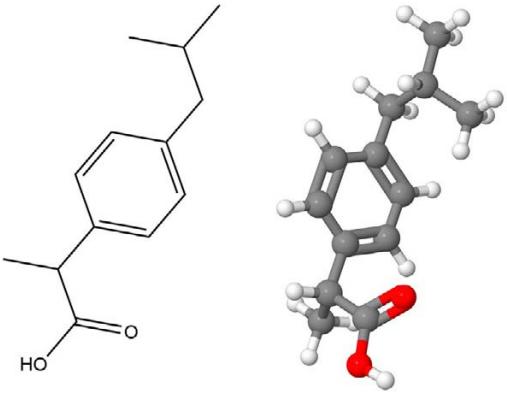
Seuraavaksi jokaista yllä mainittua vaatimusta tarkastellaan yksityiskohtaisesti.

Koneluettavat rakenteet

Suurinta osa asiakirjoista ei ole suunniteltu konekäsittelyä ajatellen, vaan ne on suunniteltu ihmisten luettavaksi. Ihmiset ovat erittäin hyviä tunnistamaan kuvioita ja prosessoimaan kieltä, mikä vaikuttaa yhdisteiden esittämiseen. Siksi on olemassa hyvin vakiintunut tekstinlouhinnan ja luonnollisen kielen käsittelyn tutkimuskenttä, joka keskittyy tietokoneella ymmärrettävän tiedon eristämiseen ihmisille suunnitelluista asiakirjoista. Asiakirjojen kemialliset yhdisteet ja niiden rakenteet vaikuttavat olevan erityisen vaikeita tunnistaa ja käsitellä tietokoneella. Mieti esimerkiksi

ibuprofeenia ja kaikkia tapoja, miten siihen voidaan viitata asiakirjassa:

- synonyymejä: ibuprofeeni, Motrin, Andran, Brufen, Liptan, Advil, Butylenin, Ibuprocen, Anflagen, Buburone, 2-[4-(2-Methylpropyl)Phenyl]Propanoic Acid[54]
- イブプロフェン (ibuprofeeni japanin kielellä)
- tekstillä "all over-the-counter NSAIDs" (ibuprofeeni on useimmissa maissa käsikauppalääkkeenä myytävä ei-steroidinen tulehduskipulääke[55]
- 2D- ja 3D-rakenteilla (jopa osana reaktioyhtälöä) (kuva 10).



Kuva 10. Ibuprofeenin 2D- ja 3D-esitykset.

Ylivoimaisesti paras tapa lähesty haastetta on, että asiakirjojen laatijat ovat tietoisia kemiallisten rakennetietojen konelukemisen tarpeesta. Rakenteet tulee toimittaa koneluettavassa standardimuodossa (SMILES, InChI, ChemDraw-tiedosto jne.) yhdessä

[54] Suomeksi (*RS*)-2-[4-(2-metyylipropyyli)fenyyli]propaanihappo

[55] Alkuperäiskielellä "over-the-counter non-steroidal anti-inflammatory drug"

artikkelin kanssa siten, että niihin on linkattu tai viitattu asiakirjassa. Ne voidaan myös lisätä artikkelin lisäaineistoksi.

Harvat verkkosivut tai tieteelliset julkaisut sisältävät kaikki nämä tiedot, mutta tilanne muuttuu nopeasti useiden lehtien merkatessa artikkeleiden sisältämiä rakenteita esimerkiksi linkittämällä ne PubChem ja ChemSpider -tietokantoihin.

Edelläkävijöitä ovat muun muassa Nature Chemistry[56] ja jotkut Prospect-projektiin sisällytetyt RSC-lehdet[57]. Huomaa, että suurin osa artikkeleista on tällä tietokoneille haastavassa muodossa. Esimerkiksi PDF-tiedostot ovat esteettisesti erittäin hyviä, mutta "tuhoavat" tietoa (esim. taulukosta tulee kuva tai pelkkää tekstiä). PDF:n vaihtoehdoksi on ehdotettu datument-asiakirjaa. [58]

Koska useimmat asiakirjat eivät tällä hetkellä tarjoa kemiallista rakenneinformaatiota, on yritettävä luoda koneellisesti luettavaa informaatiota ihmisille luodusta informaatiosta. Monimutkaisuuden vuoksi tähän ei ole täydellistä tapaa, mutta on olemassa muutamia prosessia auttavia työkaluja:

- nimiontologiat (oikeastaan vain synonyymihakuja)
- nimestä rakenteeksi -konvertointiohjelmistot (esim. OSCAR3[59] ja Lexichem[60]) (engl. *name to structure*)
- kuvasta rakenteeksi -konvertointiohjelmistot (esim. ChemReader[61] ja CLiDE).

Luonnollinen kielenkäsittely voi auttaa erottamaan rakennetiedot normaalista tekstistä syntaktisen kontekstin perusteella. Nämä menetelmät eivät ole täydellisiä, mutta toimivat melko hy-

[56] http://www.nature.com/nchem

[57] https://web.archive.org/web/20070401173200/http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp

[58] https://journals.tdl.org/jodi/index.php/jodi/article/view/130/128

[59] http://sourceforge.net/projects/oscar3-chem

[60] http://www.eyesopen.com/lexichem-tk

[61] https://doi.org/10.1186/1752-153X-3-4

vin. Haasteita on monia, kuten esimerkiksi, kuinka käsitellään viittaukset yhdisteryhmiin, geneeriset yhdisteet (esim. tulehduskipulääkkeet, COX-2-estäjät jne.) ja monimutkaiset kaaviot. Yllä esitetyssä esimerkissä on epäselvää, viittaako asiakirjan ilmaisu "all over the counter NSAIDS" ibuprofeeniin vai ei, sillä yhteys ei ole yksiselitteinen. Myös ibuprofeenin myynti käsikauppalääkkeenä saattaa muuttua ajan myötä eri maissa. Huomaa, että suositun PubMed-tietokannan artikkeleiden tiivistelmien yhdisteet on linkitetty PubChem-tietokannan yhdisteisiin yllä kuvatulla prosessilla.

Löydettävyys

Kun kemiallinen rakenneinformaatio on tunnistettu asiakirjoissa, se on helposti linkitettävissä standarditekniikoilla. Näin siitä tulee ainakin teoreettisesti löydettävää. Esimerkiksi HTML-dokumentissa voi olla yksinkertaisesti linkki rakennetiedostoon:

```
... we found that long term use of <a href="ibuprofen.sdf">Ibuprofen</a>
is associated with an elevated risk of stroke ...
```

Tai tietokantaviittaus esimerkiksi PubChemiin:

```
... we found that long term use of <a
href="http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=3672">
Ibuprofen</a> is associated with an elevated risk of stroke ...
```

Vielä parempaa olisi käyttää tag-tunnistetta XML:n tai RDF:n kanssa, vaikka tämä vaatisi termin standardoinnin. Toisin sanoen yksi asiakirja voisi käyttää ”COMPOUND”-tunnistetta, toinen voisi käyttää termiä “STRUCTURE” jne.:

```
... we found that long term use of <COMPOUND In-
ChI="InChI=1S/C13H18O2/c1-9(2)8-11-4-6-12(7-5-11)10(3)13(14)15/h4-
7,9-10H,
8H2,1-3H3,(H,14,15)">Ibuprofen</COMPOUND> is associated with an
elevated risk of stroke ...
```

Linkkien tai viitteiden lisääminen ei ratkaise hakuongelmaa, mutta teoreettisesti hakukoneet tai indeksoijat voivat paikantaa rakenneviitteet tallentaa ne hakutietokantaan. Tällä hetkellä Googlen kaltaiset hakukoneet eivät ymmärrä kemiallisia rakenteita tai mahdollista niiden hakua. Näin ollen löydettävyys riippuu siitä, pystytäänkö rakenne löytämään asiakirjasta hakukoneen tekstikyselyllä.

Verkkoympäristössä SMILES- ja InChI-kaavat ovat haastavia hakea niiden koon ja välimerkkien käytön vuoksi, mikä johtaa korkeaan väärien positiivisten havaintomäärään. Esimerkiksi SMILES-kaavojen C, CC, CCC tai jopa CC (= O) CC etsiminen ei palauta kemiallista rakennetta Googlen suosituimmissa osumissa. Tällä hetkellä ainoa tapa saavuttaa hakukonelöydettävyys on hyödyntää teemassa 2 esiteltyä InChI-avainta.

Kontekstualisointi

Kun kemiallisen rakenneviitteen sijainti asiakirjassa on tunnistettu, voidaan molekyyli kontekstualisoida tarkastelemalla sanoja, joita asiakirjan rakenteisiin viittaavat virkkeet ja kappaleet sisältävät. Tätä varten on kehitetty esimerkiksi seuraavat menetelmät: tilastollinen analyysi (yhdisteen esiintymistä yhdessä muiden kiinnostavien termien kanssa analysoidaan tilastollisesti) ja luonnollinen kielen käsittely (analysoidaan teksti syntaksin ymmärtämiseksi).

Yksinkertaisin tilastollinen analyysi etsii tekstistä termien rinnakkaispaikkoja. Esimerkiksi ei voida tarkastella pelkästään yksittäisiä yhdisteitä tekstissä, vaan kaikkia yhdisteitä ja niiden liittymistä toisiinsa. Ovatko ne samanlaisia? Onko olemassa tiettyjä ryhmiä? Yhdisteiden rinnakkaista esiintymistä voidaan analysoida

myös tarkastelemalla aiheeseen liittyvien ontologisten termien esiintymistä.

Aikaisemmassa tekstianalyysitutkimuksessa biologiassa ja muilla aloilla on tarkasteltu esimerkiksi seuraavia kysymyksiä: Kuinka tieteellisen julkaisun tiivistelmää painotetaan suhteessa kokotekstiin tai rakenteiden rinnakkaisten sijaintien suhteellista painotusta lauseessa, kappaleessa tai asiakirjassa (lisätietoa artikkeleissa Lin, 2009; McIntosh & Curran, 2009)?

Luonnollisen kielen prosessointi vie tämän pidemmälle ymmärtämällä, millaisia sanoja asiakirjassa on (esim. substantiivit, verbit, prepositiot jne.). Tämä mahdollistaa todellisten suhteiden muodostumisen (esim. yhdiste x estää proteiinia y). Jonkin verran alustavaa työtä tämän suhteen on tehty Indianan yliopistossa (Jiao & Wild, 2009).

Saavutettavuus

Saavutettavuudelle on kolme relevanttia tasoa: pääsy kemiallisiin rakenteisiin, pääsy rakenteita kontekstualisoiviin tietoihin ja pääsy artikkelin kokotekstiin. Vaikka verkkosivut yleensä sallivat pääsyn kaikille kolmelle tasolle, tieteellisissä julkaisuissa saavutettavuus vaihtelee. Tällä hetkellä yhdessäkään lehdessä ei ole täysin avointa pääsyä kemiallisiin rakenteisiin ja ontologiatietoon (ts. mahdollisuus ladata täysin ontologisesti merkittyjä artikkeleita), lukuun ottamatta edellä mainittuja rajoituksia. Open Access -lehdet sallivat ilmaisen pääsyn artikkelin kokotekstiin, mutta kemian lehtien lukumäärä on hyvin rajallinen (esim. Chemistry Central Journal[62]), ja pääsy rajoittuu yleensä HTML- ja PDF-muotoihin. Tilanne muuttuu koko ajan juuri laaditun mandaatin myötä, jonka mukaan kaikkien Yhdysvaltain valtion rahoittamien tutkimusjulkaisujen on oltava vapaasti saatavissa vuoden kuluttua. Myös rakenteilla oleva PubMed-keskusarkisto[63] voi vaikuttaa myönteisesti saavutettavuuteen.

[62] http://journal.chemistrycentral.com

[63] https://www.ncbi.nlm.nih.gov/pmc

Tehtävät

24. Mikä on 2-etyyli-2-hydroksi-3-okso-butaanihapon kiehumispiste? Yritä selvittää se hakemalla verkosta “2-ethyl-2-hydroxy-3-oxo-butanoic acid boiling point” ja “VUQLHQFKACOHNZ-UHFFFAOYSA-N boiling point”. Vertaile saatuja tuloksia.
25. Hae PubMed-sivustolla[64] termejä “paracetamol” ja “acetaminophen”. Miksi jotkut tuloksista eivät sisällä määritettyä termiä? Miksi luulet, että osumia on melkein sama määrä, mutta ei aivan?
26. Millaisia haasteita voisi olla paperilla esitetyn yhdisteen IC_{50}- biologisen aktiivisuuden kontekstualisoinnissa?

[64] https://pubmed.ncbi.nlm.nih.gov

8 KEMINFORMATIIKKA KEMIAN KIRJASTOISSA

Teeman tavoitteena on

- tutustua kaupallisiin keminformatiikkatyökaluihin ja tietoaineistoihin
- tutustua kemian kirjastoille relevantteihin ilmaisiin keminformatiikkatyökaluihin ja tietoaineistoihin
- ymmärtää kemian kirjastoissa käytettävien tiedonhankinta-työkalujen ja keminforma-tiikan nousevia trendejä.

Kaupalliset keminformatiikkaresurssit

Kemian kirjastojen kaupalliset työkalut ovat yleensä kehittyneet kaupallisten ohjelmistojen tuottajien ja myyjien välisistä kumppanuuksista, indeksointipalveluista ja joissain tapauksissa tieteellisten lehtien kustantajista. Näistä tunnetuin on Chemical Abstracts Servicen, FIZ Chemien ja Japan Science and Technology Corporationin välinen kumppanuus, josta on syntynyt useita satoja tietokantoja STN-tuotemerkin alle.[65]

Perinteisesti näitä tietokantoja käytettiin asiantuntijoiden hakutyökaluilla, mutta nykyisin niitä voi käyttää suositulla ja käyttäjäystävällisellä SciFinderilla.[66] SciFinder mahdollistaa teksti-, rakenne- ja samankaltaisuushaut kirjallisuudesta ja patenteista poimitusta kuratoidusta tietoaineistosta. Haettavissa olevat CAS-tietoaineistot sisältävät CAS-rekisterin (suuri määrä kemiallisia yhdisteitä), CAS-Reactin (kirjallisuudesta eristetyn reaktiotietoaineiston) ja MARPATin (patenteista eristettyjä geneerisiä rakenteita ja niihin liittyviä tietoja).

Elsevierin Reaxys-palvelu[67] on vaihtoehto SciFinderille. Sen kautta voi käyttää Beilstein-tietoaineistoa. Beilstein on rakennettu the Handbook of Organic Chemistry -käsikirjasta ja Gmelin-tietoaineistosta, joka on suuri organometallien ja epäorgaanisten yhdisteiden tietoaineisto.

Pienempiä kaupallisia tietoaineistoja ovat esimerkiksi GVKBIO[68] (hyväksyt lääkkeet, kliiniset ehdokkaat ja yhdisteiden inhibointikohdetiedot), WOMBAT[69] (tietoaineisto, joka linkittää yhdisteet kohteisiin ja kohteiden sekvenssitietoihin), MDDR (MDL Drug Data Report)[70] (tietoa lääkeaineista ja mahdollisista

[65] https://www.cas.org/support/training/stn

[66] https://www.cas.org/products/scifinder

[67] https://www.reaxys.com

[68] http://www.gvkbio.com

[69] https://web.archive.org/web/20090215051257/http://sunsetmolecular.com. Huom. WOMBAT on nykyisin Sigma-Aldrichin tuote.

[70] http://www.akosgmbh.eu/accelrys/databases/mddr.htm

lääkeaineista, eristetty julkaistuista asiakirjoista, kokousraporteista ja konferenssikirjoista) ja DNP (Natural Products Dictionary)[71] (yhdisteiden kemiallisia ja biologisia tietoja).

Ilmaiset keminformatiikkaresurssit

Kemian kirjastoille on tarjolla verkossa laaja valikoima ilmaisia resursseja, tietoaineistoja ja työkaluja. Yleensä niitä käytetään kaupallisten resurssien lisämateriaalina. Erinomainen metaresurssi on Gary Wigginsin kehittämä Chemical Information Sources Wiki[72], joka sisältää useita lukuja erityyppisten kemiallisten tietojen etsintästrategioista. Wikissä on myös SIRCh-aineisto[73] (Selected Internet Resources for Chemical), joka sisältää linkit tietokantojen verkkoresursseihin ja työkalut näiden hakujen suorittamiseen. Lisäksi Student's Guide to Free Chemistry Software -sivusto[74] esittelee kemian opiskelijoille hyödyllisten ohjelmistojen kokoelman.

Ei ole olemassa ilmaisia työkaluja, jotka mahdollistavat manuaalisesti parannettujen tietoaineistojen käytön (ChEMBL[75] tarjoaa tämän melkein). Silti avoimet tietokannat kuten PubChem ja ChemSpider tarjoavat valtavan määrän tietoa. Ne kattavat yhä suuremman määrän yhdisteitä, vaikka esimerkiksi Southan et al. (2009) mukaan kaupalliset tietoaineistot tarjoavat edelleen ainutlaatuista kemiaa, jota ei tällä hetkellä löydy julkisista tietokannoista. Näistä tietoaineistoista voi hakea kemiallisia rakenteita, alarakenteita, samankaltaisuuksia tai tekstejä, joilla voidaan tunnistaa muun muassa synonyymejä, kokeellisia ja ennustettuja kemiallisia ominaisuuksia, spektrituloksia, kirjallisuusviitteitä, myyjiä, biologista aktiivisuutta, määritystuloksia ja lääketieteellisiä luokitteluja.

[71] http://www.chemnetbase.com/faces/search/SimpleSearch.xhtml

[72] http://en.wikibooks.org/wiki/Chemical_Information_Sources

[73] https://en.wikibooks.org/wiki/Chemical_Information_Sources/SIRCh

[74] https://sites.google.com/site/chemistryfreeware

[75] https://www.ebi.ac.uk/chembl

Käytettävissä on myös NIST Chemistry WebBook[76], joka linkittää laajasti yhdisteitä ja niiden fysikaalisia ominaisuuksia koskevia tietoja.

Ominaisuuksien ennustamiseen on tarjolla useita ilmaisia työkaluja, jotka ovat hyödyllisiä silloin, kun kokeellista dataa ei ole. Yleensä ennustetaan ominaisuuksia kuten LogP, liukoisuus, polaarinen pinta-ala, pKa ja bioaktiivisuus. Soveltuvia ilmaisia verkkosovelluksia ovat muun muassa Molinspiration[77] ja Virtual Computational Chemistry Laboratory[78]. Lisäksi on olemassa ilmaisia työpöytäohjelmistoja, kuten MedChem Designer[79], ja Estimation Program Interface Suite[80]. Suosittuja kaupallisia työkaluja tarjoavat muun muassa ACD Labs[81] ja Simulations Plus[82].

Kemian kirjastojen nousevia trendejä

Aikaisemmat tutkimukset ovat osoittaneet, että vaikka perinteiset hakutyökalut toimivat edelleen hyvin kemisteille ja kemian informaatikoille, on geneeristen hakukoneiden suosio kemiallisten ongelmien ratkaisemisessa kasvanut erityisesti nuorten kemian tutkijoiden parissa (Banville, 2008).Vaikka yhä enemmän tietoa on saatavana sähköisesti (tieteelliset julkaisut, verkkosivut ja myös kasvava määrä kirjoja), on tiedon saavutettavuudessa vielä useita haasteita. Esimerkiksi, hakukoneet eivät indeksoi maksumuurin takana olevien tieteellisten julkaisuiden sisältöä, ja kuten teemassa 7 kuvattiin, koneellisen luettavuuden, löydettävyyden, kontekstuaalisuuden ja kemiallisen informaation saatavuuden vaatimukset täyttyvät usein huonosti.

[76] http://webbook.nist.gov/chemistry

[77] http://www.molinspiration.com/cgi-bin/properties

[78] http://www.vcclab.org

[79] https://www.simulations-plus.com/software/medchem-designer

[80] https://www.epa.gov/tsca-screening-tools

[81] http://www.acdlabs.com

[82] http://www.simulations-plus.com

Pelkästään verkkohakua koskevan lähestymistavan vaarana on, että koska verkossa on saatavana niin paljon tietoa, löytyy ”jotain tietoa” helposti, mutta samalla iso määrä tietoa jää huomiotta. Näin voi käydä etenkin kokemattomille tutkijoille, joilla on heikot tiedonhankintataidot. Akateemisessa maailmassa tiedonhankinnan nykytrendi on, että tutkijat etsivät tietoa yhä enemmän itse. Tämä on kätevää ja edullista, mutta ohittaa kemiaan erikoistuneiden informaatikoiden kokemuksen ja erikoistuneiden työkalujen käyttöosaamisen.

Nopeasti kasvavien informaatiomäärien kanssa selviytyminen merkitsee väistämättä joidenkin perinteisesti ihmisen suorittamien toimintojen automatisointia. Semanttiset teknologiat ovat osa ratkaisua, mutta automatisoinnin esteenä on vielä useita haasteita:

1. **Tieteellisten julkaisujen sisällön vapauttaminen.** Kuten edellisessä luvussa todettiin, pääsy koneellisesti luettaviin tietoihin artikkeleissa on satunnaista.
2. **Mahdollisuus hyödyntää metatietoja** laadun, luotettavuuden, kuratoinnin tason ja tiedonlähteen osoittamiseksi.
3. **Vahva turvallisuus tarvittaessa.** Turvallisuus on tärkeää sekä yliopistoissa että teollisuudessa. Erityisen tärkeää se on teollisuudessa, missä tietoturvallisuuden pettäminen voi aiheuttaa erittäin kalliita kilpailuedun menetyksiä. Akateemisessa maailmassa seuraukset ovat lievempiä, vaikka suurin osa tutkijoista haluaa suojella omaa immateriaalioikeuttaan, etenkin hauraassa tutkimuksen alkuvaiheessa.
4. **Avoin laboratoriokulttuuri mahdollisuuksien mukaan**. Tietoturvallisuuteen liittyvät haasteet ovat ensisijaisesti teknisiä, mutta tämä on enemmänkin kulttuurinen este. Erityisesti akateemiset laboratoriot tuottavat suuria määriä hyödyllistä informaatiota, mutta niitä ei koskaan julkaista, tai ne julkaistaan viiveellä. Tämä johtuu monista syistä. Esimerkiksi tulokset saattavat olla negatiivisia ja niitä ei pidetä hyödyllisenä tutkimushankkeelle. Kyse voi olla immateriaalioikeuksista tai tuloksia ei pidetä sellaisina, että ne voisi julkaista tieteellisessä artikkelissa. Julkaisemisen puolueellisuusongelmaa, eli taipumusta julkaista vain positiivisia tuloksia, on tutkittu laajas-

ti, mutta kaikki siihen kehitetyt potentiaaliset ratkaisut sijoittuvat keminformatiikan tutkimusalan ulkopuolelle. Immateriaalioikeuksien haaste on, että ne edellyttävät todennäköisesti tietojen yksityisyyttä.

Tehtävät

27. Työskentelet kemistin kanssa etsien tietoa tietystä korvautumisreaktiosta. Mitä kaupallista työkalua käyttäisit? Millaisen haun suorittaisit? Täydentäisitkö tiedonhankintaa ilmaisilla tai verkkopohjaisilla työkaluilla? Jos kyllä, niin millä.
28. Valitse yhdiste ja etsi sen kiehumispiste useista lähteistä – mieluiten sekä kaupallisista että ilmaisista. Vastaavatko eri lähteiden tulokset toisiaan? Kuinka päätät, mikä niistä on “oikea” vastaus?
29. Miten voisit auttaa henkilöä, joka etsii sellaisen yhdisteen LogP-arvoa, jota ei löydy julkisista tai kaupallisista tietoaineistoista?

9 KEMIALLISEN TIETOAINEISTON ANALYSOINTI KLUSTEROINNIN JA MONIMUOTOISUUDEN AVULLA

Teeman tavoitteena on ymmärtää

- klusterianalyysin kolme keskeisintä keminformatiikan sovellusta
- hierarkkisten klusterointi-menetelmien toiminta
- epähierarkkisen ja hierarkkisen klusteroinnin erot
- deskriptoriavaruuden, kemiallisen avaruuden ja lääkeaineavaruuden eroavaisuudet
- peittomenetelmien ja suhteellisen monimuotoisuuden väliset eroavaisuudet.

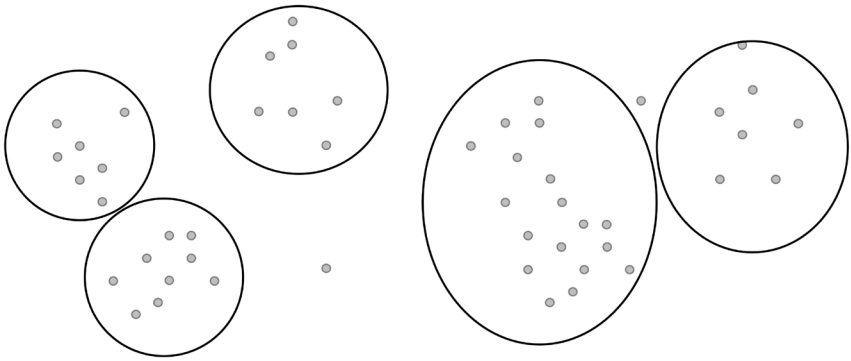
Johdanto

2D- ja 3D-rakenteiden samankaltaisuusmittaukset ja deskriptorit mahdollistavat yhdessä tietopisteiden järjestämiseen ryhmiin samankaltaisuuden perusteella (klusterianalyysi), tai niiden keskinäisten samankaltaisuuksien ja eroavaisuuksien analysoinnin (monimuotoisuusanalyysi).

Klusterianalyysi

Klusterianalyysillä tarkoitetaan tilastollisia menetelmiä, joita käytetään samankaltaisten ryhmien (klusterit) tunnistamiseen moniulotteisessa avaruudessa. Tämä edellyttää määritettävien kohteiden samankaltaisuusanalyysiä. Menetelmää voidaan havainnollistaa ajattelemalla yksinkertaista kaksiulotteista avaruutta, jossa visualisoidaan kahta ominaisuusdeskriptoria.

Kuvassa 11 harmaat pienet ympyrät edustavat avaruuteen piirrettyjä yhdisteitä, ja mustan ääriviivan omaavat ympyrät mahdollisia klustereita. Huomaa, että klusterointi ei ole ehdotonta. Ryhmien määrittäminen on subjektiivista riippuen niiden tunnistamiseen käytetystä menetelmästä. Esimerkiksi kuvan kaksi vasenta klusteria voisivat olla myös yksi klusteri. Klustereiden ulkopuolisia yksittäisiä datapisteitä kutsutaan singletoneiksi.



Kuva 11. Klusterianalyysin esimerkkivisualisointi.

Klusterianalyysimenetelmiä on sovellettu monilla eri alueilla, ja niitä käytetään laajalti esimerkiksi tiedonlouhinnassa, hahmontunnistuksessa ja koneoppimisessa. Keminformatiikassa klusterointia käytetään kolmeen päätarkoitukseen:

- **Yhdisteiden ryhmittely kemiallisiin sarjoihin** (tai joihinkin sarjoihin rinnastettaviin) on tapa organisoida laajat tietoaineistot. Kemistin on helpompi selata 500 klusteria, joka sisältää samanlaisia molekyylejä, kuin 50 000 mielivaltaisesti järjestettyä yhdistettä.
- **Uusien bioaktiivisten molekyylien tunnistaminen:** Jos bioaktiivisuuden näkökulmasta tuntematon yhdiste löytyy klusterista, jonka yhdisteillä on tunnettu aktiivisuus, voidaan tuntemattoman yhdisteen aktiivisuudesta ennustaa todennäköisyys. Esimerkiksi, jos klusterin yhdisteistä 75 % on aktiivisia, voidaan tuntemattoman yhdisteen olevan bioaktiivinen 75 % todennäköisyydellä.
- **Edustavien alaryhmien valitseminen**: Kun molekyyliaineisto klusteroidaan, voidaan jokaisesta klusterista valita yksi yhdiste edustamaan klusteria. Valitut yhdisteet edustavat koko tietoaineistoa, mikä on joskus hyödyllisempää kuin satunnainen valinta.

Hierarkkinen klusterointi

Klusterointimenetelmät voivat olla hierarkkisia tai epähierarkkisia. Hierarkkinen klusterointi luo puumaisen klusterirakenteen, jonka alimmalla tasolla jokainen alkio on omassa klusterissaan ja ylimmällä kaikki alkiot yhdessä klusterissa. Algoritmisesti tämä voidaan tehdä joko aloittamalla alhaalta klustereita yhdistäen (engl. *agglomerative*) tai ylhäältä klustereita jakaen (engl. *divisive*).

Useimmiten agglomeratiiviset (tai kokoavat) hierarkkiset menetelmät toimivat algoritmisesti aina samalla tavalla, mutta eroavat toisistaan siinä suhteessa, miten klustereiden yhteen sulauttaminen kullakin tasolla päätetään. Esimerkiksi Wardin menetelmää

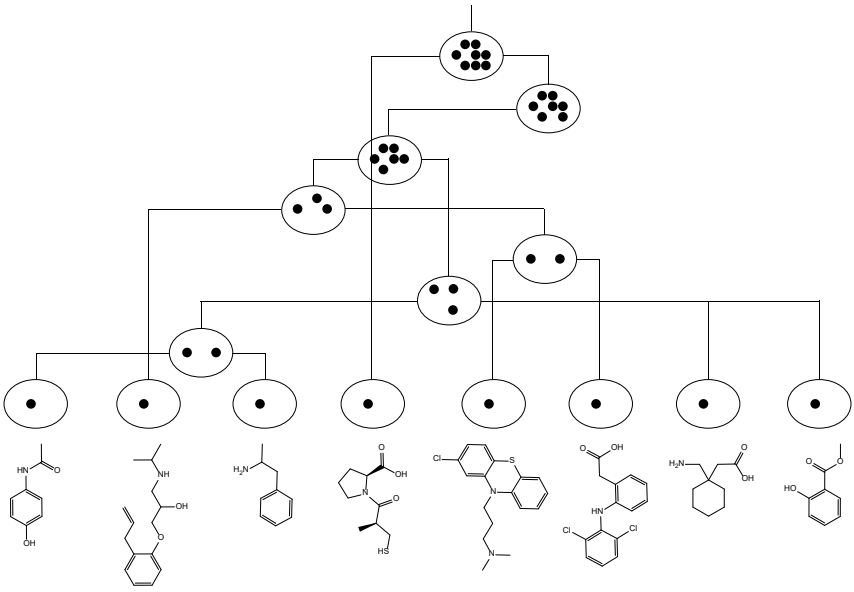
on käytetty keminformatiikassa laajasti (Wild & Blankley, 2000). Siinä klusterit yhdistetään siten, että sulautettujen klusterien varianssin keskiarvo kasvaa mahdollisimman vähän eli luodaan "tiukin" mahdollinen klusteri. Muita menetelmiä ovat yksittäinen kytkentä[83] (klusterit yhdistetään kunkin klusterin lähimpien pisteiden minietäisyyden mukaan), täydellinen kytkentä (klusterit yhdistetään kunkin klusterin kauimpien pisteiden minietäisyyden mukaan) ja ryhmäkeskiarvo (kahden klusterin kaikkien parien minimi etäisyyskeskiarvo).

Hierarkkiset menetelmät ovat laskennallisesti monimutkaisia ja vaativat paljon muistia, joten ne eivät skaalaudu hyvin suurten tietoaineistojen prosessointiin. Ne voidaan klusteroida nopeammin epähierarkkisilla menetelmillä. Tietojoukon ositetun ryhmittelyn luomiseksi (ts. jokainen alkio on vain ja ainoastaan yhdessä klusterissa tai on singleton), on klusteripuusta valittava horisontaalinen vaakaotos. Ainoa laajasti käytetty jakava menetelmä, on nimeltään Divisive *K*-means -menetelmä.

Tutkitaan näiden menetelmien toimintaa tarkastelemalla tietoaineistoa, jota käytettiin aikaisemmin 2D-tietokantaesimerkissä (ks. luku 4). Hierarkkiset menetelmät asettavat aluksi jokaisen datapisteen itsenäiseksi klusteriksi alatasolle (*n* klusteria, joissa *n* = pisteiden lukumäärä). Seuraavaksi tunnistetaan kaksi klusteria, jotka sulautetaan yhdeksi klusteriksi seuraavalle tasolle. Tunnistamislogiikka riippuu käytetystä menetelmästä, kuten yllä kuvattiin. Seuraava taso ylöspäin muodostuu siis yhdestä klusterista, jossa on kaksi pistettä, ja kaikista muista yhden pisteen klustereista. Toisin sanoen siellä on *n*–1 klusteria. Prosessi toistuu, kunnes hierarkian yläosassa on vain yksi klusteri sisältäen kaikki pisteet. Näin muodostuu klusterihierarkia (ks. kuva 12). Hierarkiasta voidaan erottaa osio (esim. yhdisteiden ryhmä, jonka jokainen yhdiste on samassa klusterissa) valitsemalla hierarkiasta haluttu "taso".

[83] Kytkennän alkuperäistermi englanniksi: *linkage*

Hyvien tasojen valitsemiseksi on olemassa monia algoritmeja (ks. Wild & Blankley, 2000)[84].



Kuva 12. Esimerkki hierarkkisen klusteroinnin etenemisestä.

Epähierarkkinen klusterianalyysi

Epähierarkkisissa menetelmissä voidaan käyttää monenlaisia algoritmeja, mutta yleensä ne klusteroivat tietoaineiston yhdellä osioinnilla (verrattuna puuhun, joka voi johtaa moniin osioihin). Tunnettuja esimerkkejä epähierarkkisista menetelmistä ovat Jarvis-Patrick, *K*-means ja *k*-medoids.

Jarvis-Patrick (JP) on epähierarkkinen menetelmä, jossa jokaiselle ryhmän yhdisteelle tunnistetaan lähimmät naapurit (*j*) (ts. tietoaineiston muut yhdisteet, jotka ovat kaikkein samankaltaisimmat).

[84] https://doi.org/10.1021/ci990086j

Yhdisteet sijoitetaan samaan klusteriin, jos ne a) löytyvät toistensa samankaltaisuuslistasta *j* ja b) niillä on sama *k* kuin lähimmällä naapurilla *j*. Tämä menetelmä ei vaadi tason valintaa, mutta edellyttää *j*:n ja *k*:n ennalta määrittämistä. Tanimotoa käytetään yleensä samankaltaisuuden mittana, sillä vaikka JP on nopea, ovat sen tulokset keminformatiikassa ovat olleet laadultaan vaihtelevia.

K-means-menetelmää käytetään laajemmin kuin JP-menetelmää. Se edellyttää, että haluttujen klustereiden lukumäärä *m* tiedetään etukäteen. Alkuperäinen joukko *m*-klusterin painopisteitä luodaan esimerkiksi valitsemalla sattumanvaraisesti yhdisteet, joita käytetään keskipisteinä. Jokainen nimike *n* sijoitetaan lähimpään klusteriin laskemalla kohteen ja kunkin klusterin keskikohdan samankaltaisuus. Kun kaikki kohteet on analysoitu, lasketaan vasta muodostetuille klustereille uudet keskipisteet. Koska tämä muuttaa klusterien määrityksiä, prosessia toistetaan määrittämällä klusterien jäseniä uudelleen niin kauan, että klusterit eivät enää muutu (ts. klusterit ovat vakaita). Yleensä tarvitaan vain muutamia iteraatioita (<100, usein <10).

k-medioids on *K*-means-menetelmän johdannainen, joka toteutetaan PAM-menetelmänä (Partitioning Around Medioids) R-ohjelmointikielessä, ja käytetään siksi yleisesti tässä ympäristössä. Se eroaa *K*-means-menetelmästä siten, että se käyttää todellisia esimerkkejä tai "medioideja" klusterikeskusten kuvaamiseksi painopisteiden sijaan.

Klusteroinnista keminformatiikassa on julkaistu useita tieteellisiä artikkeleita. Hyviä johdatuksia aiheeseen ovat:

- Willett, P., Barnard, J. M., & Downs, G. M. (1998). Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences, 38*(6), 983–996. https://doi.org/10.1021/ci9800211
- Barnard, J. M., & Downs, G. M. (1992). Clustering of chemical structures on the basis of two-dimensional similarity measures. *Journal of Chemical Information and Computer Sciences*, *32*(6), 644–649. https://doi.org/10.1021/ci00010a010

- Downs, G. M., & Barnard, J. M. (2003). Clustering Methods and Their Uses in Computational Chemistry. Teoksessa K. Lipkowitz & D. Boyd (toim.), *Reviews in Computational Chemistry* (ss. 1–40). John Wiley & Sons, Ltd. https://doi.org/10.1002/0471433519.ch1

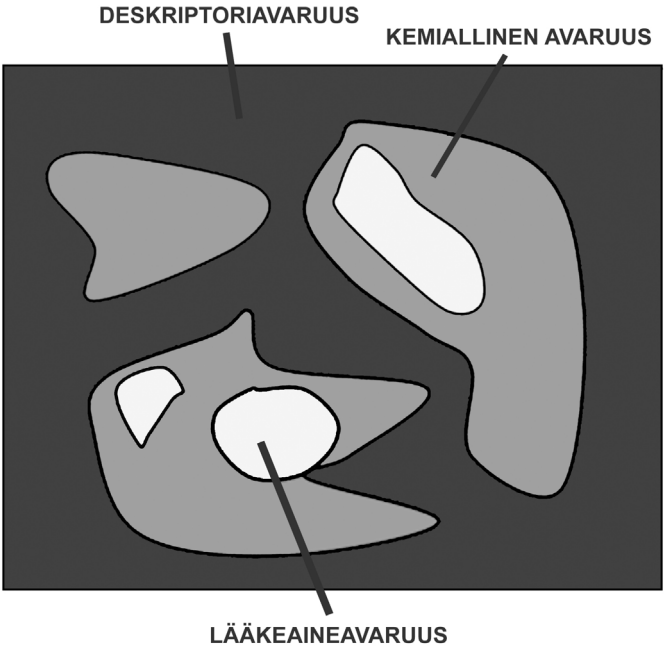
Monimuotoisuusanalyysi

Monimuotoisuusanalyysi saavutti suosiota 1990-luvun lopulla vastauksena seuraaviin lääketeollisuuden tarpeisiin:

- Lääkeyhtiöitä kiinnosti selvittää, kuinka kattavia yritysten hallinnoimat yhdistetietokannat ovat.
- Kombinatoriaalinen kemian tutkimus tuotti monia uusia yhdisteitä, ja yritykset halusivat tietää, lisäisivätkö uudet yhdisteet niiden tietokantoihin jotain uutta kemiallisen tai biologisen funktionaalisuuden näkökulmasta. Toisin sanoen, tekisivätkö uudet yhdisteet tietokannoista monimuotoisempia vai toistivatko ne niiden sisältämää informaatiota.
- Markkinoille tuli uusia kaupallisia tuhansien yhdisteiden laajuisia kirjastoja. Olivatko ne rahan arvoisia?

Tämä aloitti keskustelun "deskriptoriavaruudesta" (engl. *descriptor space*) eli moniulotteisesta euklidista avaruudesta, joka on luotu käsittelemällä yhdisteen kutakin deskriptoria yhtenä ulottuvuutena.

Erityisen mielenkiintoista oli se, kuinka deskriptoriavaruus voisi kuvata käsitteellistä kemiallista avaruutta. Esimerkiksi, jos kaikki teoreettisesti valmistettavissa olevat yhdisteet olisi jo tehty, niin kemiallinen avaruus muodostuisi moniulotteisessa deskriptoriavaruudessa alueesta, jossa määritetty deskriptorijoukko olisi edustettuna. Asetelma olisi sama myös lääkeaineavaruuden tapauksessa, joka muodostuisi lääkeainemolekyylejä sisältävästä kemiallisen avaruuden osasta. Ilmiötä voidaan havainnollistaa käsitteellisesti kuvitteellisella kaksiulotteiselle avaruudella (ks. kuva 13).



Kuva 13. Deskriptoriavaruuden, kemiallisen avaruuden ja lääkeaineavaruuden välinen hierarkia.

Kattavuus ja solupohjaiset menetelmät

Kun avaruuden eri alueet on määritetty, voidaan analysoida, mitä kemiallisen tai lääkeaineavaruuden osia kokoelmasta ei löydy. Näin tietokantaan voidaan lisätä yhdisteitä, jotka monipuolistavat kokoelmaa tehden siitä kattavamman. Esimerkiksi, jos halutaan tunnistaa lääkeaineet kemiallisesta avaruudesta, niin suoraviivaisin tapa olisi valita kaksi tai kolme deskriptoria ja piirtää yhdisteet niistä luotuun kaksi- tai kolmiulotteiseen avaruuteen. Tämä osoittaisi, ovatko yhdisteet lääkeaineita vai eivät (ks. esimerkki artikkelista Shemetulskis et al. (1995)[85]).

[85] https://doi.org/10.1007/BF00123998

Kattavuutta voidaan analysoida myös solupohjaisilla menetelmillä. Ne diskretisoivat jokaisen ulottuvuuden useaksi säiliöksi, ja luovat euklidiseen avaruuteen soluja säiliöiden leikkauskohtiin. Solujen täyttöasteen avulla voidaan määrittää, onko jossain päin suhteellista yli- tai aliedustusta koko avaruuteen verrattuna.

Suhteellinen monimuotoisuus

Suhteellisen monimuotoisuuden mittaamiseksi on kehitetty useita menetelmiä. Toisin sanoen, kuinka sisäisesti erilaisia tietojoukon yhdisteet ovat. Suoraviivaisin tapa tämän mittaamiseen on laskea tietojoukon kaikkien yhdisteparien keskimääräinen molekyylin sisäinen samankaltaisuus käyttäen Tanimoto-kerrointa, ja vähentämällä tulos arvosta 1 monimuotoisuuden mittana. On tärkeää huomata, ettei tämä kerro mitään tietojoukon kattavuudesta, vaan siitä, kuinka erilaisia sen yhdisteet ovat toisiinsa verrattuna.

Tietojoukolle voidaan myös suorittaa klusterianalyysi ja tunnistaa, mitkä klusterit liittyvät mielenkiinnon kohteena olevien yhdisteiden (esim. lääkeainemolekyylien) korkeampaan tai matalampaan esiintyvyyteen suhteessa koko aineistoon.

Tietojoukkojen vertailu

Kattavuutta ja suhteellisia menetelmiä voidaan käyttää sekä tietojoukkojen vertailuun että yksittäisten aineistojen tutkimiseen. Esimerkiksi kysymykseen ”Kuinka monipuolinen joukko A on verrattuna joukkoon B?” voitaisiin vastata vertaamalla joukkojen kattavuutta (esim. täytettyjen solujen lukumäärä) tai vertailemalla niiden keskimääräisiä eroja. Kysymykseen "Kuinka erilaisia nämä kaksi tietojoukkoa ovat?" voidaan vastata tarkastelemalla niiden päällekkäisyyttä peittomenetelmillä tai analysoimalla, kuinka yhden tietojoukon keskimääräinen eroavuus muuttuu lisäämällä analyysiin toinen tietojoukko.

Monimuotoisen osajoukon valinta

Usein on toivottavaa pystyä erottamaan monimuotoinen osajoukko. Monimuotoinen osajoukko on yhdisteiden osajoukko tietokannassa, joka edustaa koko tietojoukon kemiallista tai biologista monimuotoisuutta. Huomaa, että tämä eroaa satunnaisen osajoukon ottamisesta, jossa tarkoituksena on ottaa näyte tietojoukon koko jakaumasta.

Monimuotoisen osajoukon valintaan on useita tapoja. Solupohjaisella peittomenetelmällä voidaan valita yksi yhdiste edustamaan koko solua tai solujen joukkoa, jos täyttyneitä soluja on enemmän kuin halutun alaryhmän jäseniä. Suhteellisessa lähestymistavassa sovelletaan molekyylien sisäistä keskimääräistä samankaltaisuutta erilaisuuteen perustuvaan yhdistevalintaan (engl. *Dissimilarity-based Compound Selection* (DBCS)). Esimerkiksi valitaan ensimmäinen yhdiste satunnaisesti, ja seuraava yhdiste siten, että se on mahdollisimman erilainen kuin ensimmäinen. Kolmas valitaan siten, että mahdollisimman erilainen kuin kaksi ensimmäistä jne.

Ongelman ratkaisemiseen voidaan soveltaa myös klusterianalyysiä ryhmittelemällä tietojoukko *n* klusteriksi (missä *n* on halutun osajoukon koko), ja valitsemalla sitten edustaja jokaisesta klusterista. Edustaja voi olla esimerkiksi yhdiste, joka on lähimpänä klusterin keskiötä.

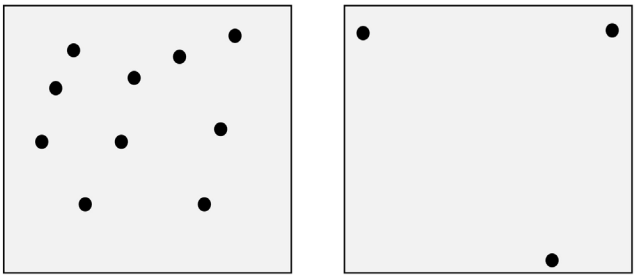
Monimuotoisuudesta on julkaistu useita artikkeleita. Alle on listattu muutamia esimerkkejä:

- Ashton, M. J., Jaye, M. C., & Mason, J. S. (1996). New perspectives in lead generation II: Evaluating molecular diversity. *Drug Discovery Today*, *1*(2), 71–78. https://doi.org/10.1016/1359-6446(96)89091-X
- Bayada, D. M., Hamersma, H., & van Geerestein, V. J. (1999). Molecular Diversity and Representativity in Chemical Databases. *Journal of Chemical Information and Computer Sciences*, *39*(1), 1–10. https://doi.org/10.1021/ci980109e

- Turner, D. B., Tyrrell, S. M., & Willett, P. (1997). Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *Journal of Chemical Information and Computer Sciences*, *37*(1), 18–22. https://doi.org/10.1021/ci960463h
- Martin, Y. C. (1996). Challenges and prospects for computational aids to molecular diversity. *Perspectives in Drug Discovery and Design*, *7*(1), 159–172. https://doi.org/10.1007/BF03380186
- Brown, R. D. (1996). Descriptors for diversity analysis. *Perspectives in Drug Discovery and Design*, *7*(1), 31–49. https://doi.org/10.1007/BF03380180
- Dunbar, J. B. (1996). Cluster-based selection. *Perspectives in Drug Discovery and Design*, *7*(1), 51–63. https://doi.org/10.1007/BF03380181
- Shemetulskis, N. E., Dunbar, J. B., Dunbar, B. W., Moreland, D. W., & Humblet, C. (1995). Enhancing the diversity of a corporate database using chemical database clustering and analysis. *Journal of Computer-Aided Molecular Design*, *9*(5), 407–416. https://doi.org/10.1007/BF00123998

Tehtävät

30. Lukeutuuko klusterointi (ja yhdisteiden erottaminen klustereista) monimuotoisten osajoukkojen tunnistamismenetelmänä peitto-menetelmiin vai suhteellisiin menetelmiin?
31. Miten voidaan luoda korkealaatuinen hierarkkinen klusterointi tietojoukolle, joka on liian suuri Wardin menetelmälle?
32. Miksi klusterointimenetelmiä on vaikea arvioida sen suhteen, miten hyvin ne järjestävät yhdisteet kemiallisiksi sarjoiksi?
33. Kumpi alla olevista tietojoukoista on kuvitteellisen 2D-deskriptoriavaruuden mukaan monimuotoisempi?

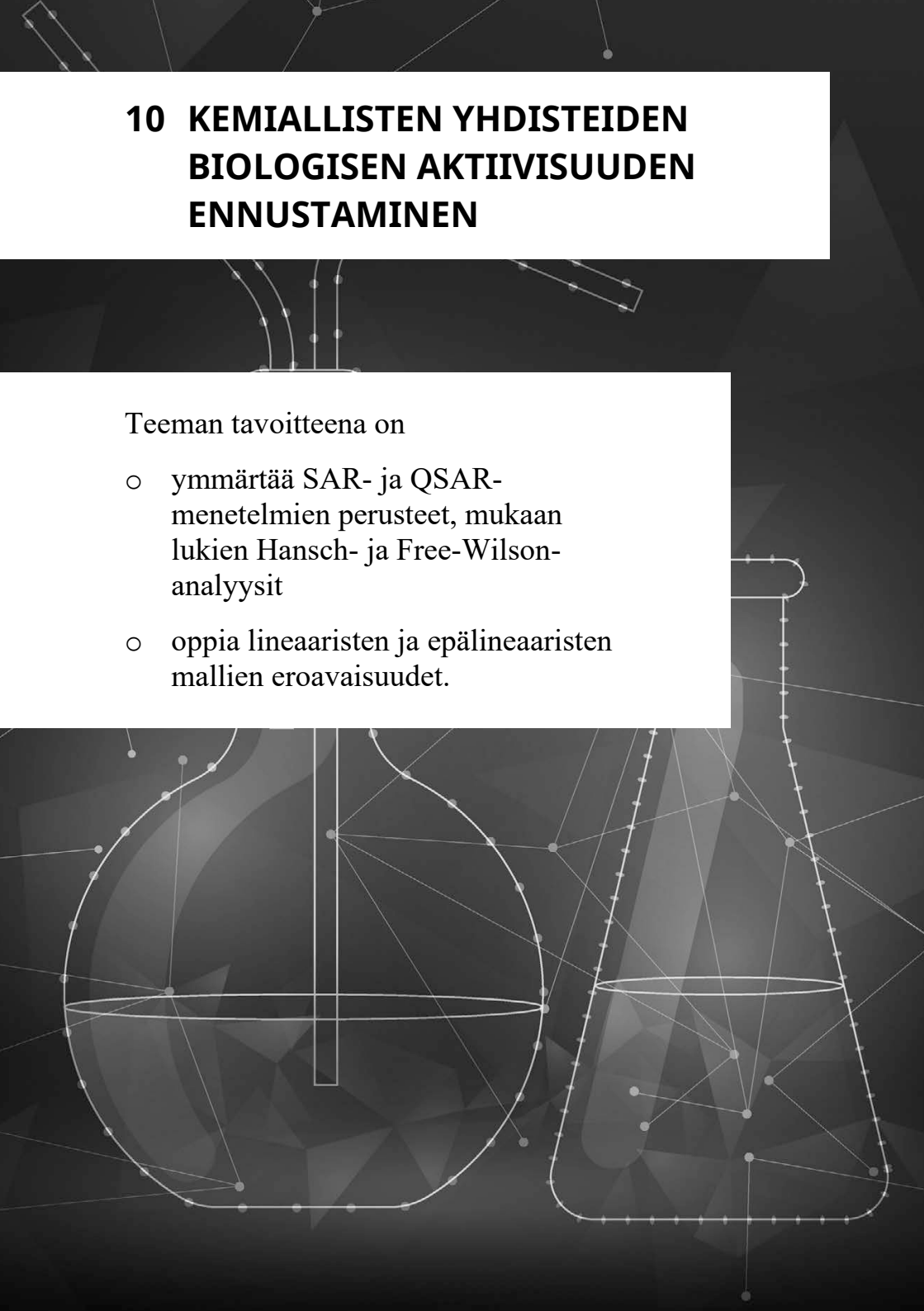


Kuva 14. 2D-deskriptoriavaruusmalleja.

10 KEMIALLISTEN YHDISTEIDEN BIOLOGISEN AKTIIVISUUDEN ENNUSTAMINEN

Teeman tavoitteena on

- ymmärtää SAR- ja QSAR-menetelmien perusteet, mukaan lukien Hansch- ja Free-Wilson-analyysit
- oppia lineaaristen ja epälineaaristen mallien eroavaisuudet.



Johdanto

Aikaisemmissa teemoissa on käsitelty 2D- ja 3D-deskriptoreiden käyttöä yhdisteiden karakterisoimisessa sekä niiden soveltavista samankaltaisuuslaskennassa, klusteroinnissa, monimuotoisuuden määrittämisessä jne. Tässä teemassa tutustutaan, miten deskriptoreita korreloidaan tuloksiin, kuten biologiseen aktiivisuuteen, kemiallisiin ominaisuuksiin ja myrkyllisyyteen. Lisäksi rakennetaan deskriptoreihin ja korrelaatioihin pohjautuvia ennustemalleja. Tämä on erittäin laaja aihe, johon tässä oppaassa perehdytään vain lyhyen johdatuksen verran.

Kvantitatiivinen rakenne-aktiivisuussuhde (QSAR)

Lääkeainekemiassa alettiin hyödyntää rakenne-aktiivisuussuhdemenetelmää (SAR) jo ennen tietokoneiden käyttöönottoa. SAR perustuu yhdisteiden rakenteellisten ominaisuuksien ja kokeellisten tulosten (usein saman sarjan) korreloimiseen. Lääkeainekemiassa syntetisoidaan usein useita lähes samankaltaisia yhdisteitä (esim. metyyli-, etyyli- ja butyylimuodot), joiden kautta tutkitaan rakenteen vaikutusta tiettyyn ominaisuuteen tai biologiseen aktiivisuuteen. Voidaan esimerkiksi havaita, että sivuketjun pidentäminen vähentää tiettyä aktiivisuutta. Rakenteen ja aktiivisuuden välistä suhdetta voidaan myös tarkastella kvantitatiivisesti.

Kvantitatiivinen rakenne-aktiivisuussuhde (QSAR) kehitettiin alun perin tavoitteena vahvistaa SAR-menetelmän matemaattista pohjaa. Erityisesti oli tarvetta pystyä määrittämään aktiivisuus tietyn deskriptorin funktiona. Deskriptoreita ja tiettyä aktiivisuutta yhdistävää funktiota voidaan soveltaa ennakoivasti aktiivisuudeltaan tuntemattomiin yhdisteisiin, jos niiden deskriptorit voidaan laskea. Huomaa, että jos aktiivisuus on ominaisuus tai myrkyllisyys, viitataan siihen tosinaan lyhenteillä QSPR tai QSTR.

Varhaisimmat QSAR-menetelmät (Hansch- ja Free-Wilson-analyysit) olivat lineaarisen regressioanalyysin sovelluksia. Hansch-analyysi liittyy ominaisuusdeskriptoreihin, ja tässä op-

paassa tarkastelunkohteena oleva Free-Wilson-analyysi rakennedeskriptoreihin. Free-Wilsonissa määritetään funktio, joka ilmaisee aktiivisuuden (määritelty log 1 / konsentraatio) painotettuina deskriptoreina, painotuksina tai kertoimina, jotka määritetään lineaarisella regressiolla. Näin saadaan yhtälö:

$$\log (1/C) = a1x1 + a2x2 + a3x3 \ldots$$

, jossa C on aktiivisuudelle vaadittava pitoisuus, $x1$, $x2$, $x3$ jne. ovat deskriptoreita (yleensä arvo 1 edustaa ominaisuuden läsnäoloa ja 0 puuttumista) ja $a1$, $a2$, $a3$ jne. ovat lineaarisesta regressiosta johdettuja kertoimia. Lineaarinen regressio on yleismenetelmä, jonka tarkoituksena on optimoida riippumattomiin muuttujiin sovellettuja kertoimia siten, että riippuvainen muuttuja (tässä tapauksessa log (1/C)) vastaa parhaiten deskriptorijoukon havaittua arvoa. Täten regressioyhtälö voidaan laatia käyttämällä tunnettujen riippuvien muuttujien arvoja, joita voidaan sitten soveltaa tuntemattomien riippumattomien arvojen ennustamiseen. Lineaarinen regressio toimii siten, että se minimoi yhtälön ennustamien arvojen ja todellisten havaittujen arvojen erotusten neliöiden summan.[86]

Jos regressioyhtälöä on tarkoitus käyttää ennakoivasti, tarvitaan jokin tapa sen tarkkuuden mittaamiseen. Yksinkertaisin tapa tarkkuuden mittaamiseen on korrelaatiokertoimen neliö (r-neliö tai r^2), joka on regressioyhtälöllä selitetty riippuvan muuttujan varianssi. Toisin sanoen, jos $r^2 = 1{,}0$; niin kaikki todelliset pisteet sijaitsevat regressioviivalla, ja jos $r^2 = 0{,}0$; niin varianssi regressioviivan ympärillä on yhtä suuri kuin riippuvan muuttujan kokonaisvarianssi. Korrelaatiokertoimen neliön haaste tarkkuusanalyysissä on, että samaa dataa käytetään sekä yhtälön rakentamiseen että tarkkuuden arviointiin. Haaste voidaan ratkaista käyttämällä q-neliötä (q^2) (kutsutaan joskus ristivalidoiduksi r-neliöksi). Täl-

[86] Suomentajan vinkki: Voit tutkia mallia tarkemmin PhET:n Least-Squares Regression -simulaatiolla (https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html).

löin yhtälöstä tehdään *n* versiota, joista jokainen versio jättää yhden alkuperäisistä tunnetuista arvoista pois (esimerkki yksi-pois-validoinnista (engl. *leave-one-out validation*). Tällöin $q2$ on keskimääräinen kokonaisvarianssi, kun yhtälöä käytetään pois jätettyjen arvojen ennustamiseen – q^2 on siten aina pienempi kuin r^2.

QSAR-menetelmällä mallinnetut biologisen aktiivisuuden tyypit sisältävät prosentuaalisen inhiboinnin. Kuinka suuren osan proteiini- tai solunäytteestä yhdiste inhiboi? IC_{50}-arvolla viitataan yhdisteen määrään, jota tarvitaan estämään 50 % näytteestä, jotka on saatu useista testeistä. Toisinaan yhdisteet luokitellaan kategorioihin ”voimakkaasti aktiivinen", "aktiivinen", "inaktiivinen" jne. Koeasetelman mukaan tulee testin virhesuhde (engl. *error rate*) ottaa huomioon.

Epälineaariset QSAR-menetelmät

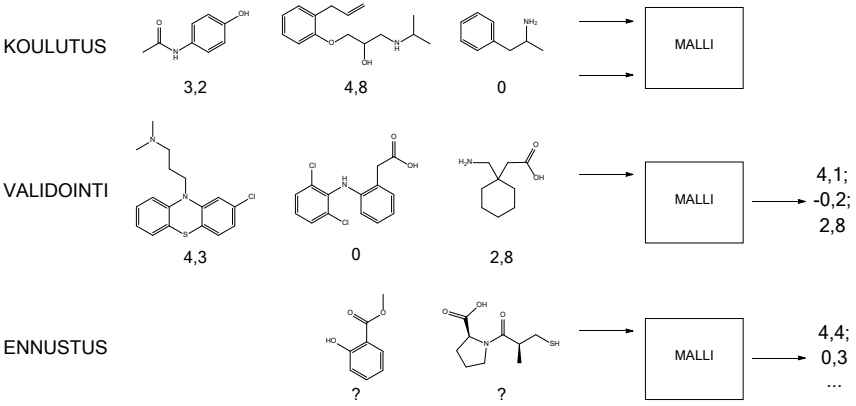
Varhaisten lähestymistapojen suurin haaste on, että niissä oletetaan aktiivisuuden vaihtelevan lineaarisesti siihen vaikuttavien deskriptoriarvojen kanssa. Yleensä tämä ei kuitenkaan ole tilanne. Epälineaariset lähestymistavat pyrkivät korreloimaan tuloksia deskriptoreihin, mutta eivät tee olettamusta lineaarisuudesta. Ne ovat siis ainakin teoreettisesti hyödyllisempiä, vaikka yleensä tehdään jonkin verran kompromisseja, esimerkiksi nopeuden, skaalautuvuuden tai tulkittavuuden suhteen. Epälineaariset lähestymistavat ovat yleensä esimerkki koneoppimisesta, erityisesti valvotusta oppimisesta. Klusterointi olisi esimerkki valvomattomista menetelmistä.

Valvotussa koneoppimisessa voidaan hyödyntää myös valvomattomien menetelmien työkaluja, kuten esimerkiksi itseorganisoituvia karttoja. Käytetty menetelmä riippuu joskus myös määritettävän QSAR-menetelmän tyypistä. Erityisesti luokittelujen yhteydessä (Ovatko yhdisteet aktiivisia vai inaktiivisia?) ja kvantitatiivisten ennusteiden laadinnassa (kun halutaan ennustaa aktiivisuusarvo). Yleisimpiä epälineaarisia menetelmiä QSAR:lle ovat:

- neuraaliverkot (engl. *neural networks*)[87]
- päätöspuut (engl. *decision trees*)[88], kuten rekursiivinen osioinnti ja satunnaismetsä (engl. *random forest*)[89]
- tukivektorikone[90] (engl. *support vector machine*)
- bayesianin-mallinnukset[91].

Eri menetelmillä on erilaiset vahvuudet ja heikkoudet. Esimerkiksi neuraaliverkot ovat "mustan laatikon" lähestymistapa, joten ne eivät ole hyödyllisiä, jos halutaan tietää, miksi tietty ennuste tehtiin. Päätöspuita taas voidaan käyttää vain luokitteluun.

Menetelmästä riippumatta mallin rakentaminen tapahtuu yleensä kolmessa vaiheessa: **koulutus** (mallin rakentamiseksi tunnetun datan pohjalta), **validointi** (mallin testaaminen tunnetulla datalla (esim. validointijoukko), jota ei käytetty mallin rakentamisessa) ja **ennustus** (mallin käyttäminen tuntemattomaan dataan).



Kuva 14. Havainnekuva valvotusta koneoppimisesta.

[87] https://en.wikipedia.org/wiki/Neural_network

[88] https://en.wikipedia.org/wiki/Decision_tree_learning

[89] https://en.wikipedia.org/wiki/Random_forest

[90] https://fi.wikipedia.org/wiki/Tukivektorikone

[91] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

Kun otetaan huomioon kemiallisten deskriptoreiden lähde (esim. sormenjäljet, ominaisuusarvot jne.), ja näiden deskriptoreiden avulla esitetyt yhdisteiden riippuvat muuttujat (kuten biologinen aktiivisuus määrityksissä), niin mallien rakentamiselle on tarjolla useita geneerisiä työkaluja. Niitä voidaan rakentaa ilmaisilla tilasto- ja tiedonlouhintaohjelmistoilla (esim. R[92] ja WEKA[93]) ja erityisesti beyesian-malleja Pipeline Pilotilla[94]. Lisäksi myös Knime-alustaa käytetään myös laajasti.[95]

Virtuaaliseulonta

Virtuaaliseulonta on yksi todella yleinen QSAR-mallien sovellus – erityisesti epälineaaristen mallien kohdalla. Tämä tarkoittaa tietokonemallien käyttöä sen sellaisten yhdisteiden ennustamiseen, jotka sitoutuvat proteiinikohteeseen tai inhiboivat solumäärityksen. Se on siten laskennallinen vastine suuritehoiselle biologiselle määritykselle. QSAR-mallien lisäksi virtuaaliseulontaan voidaan käyttää muitakin menetelmiä, kuten esimerkiksi kemiallisen samankaltaisuuden vertaamista tunnettuihin aktiivisiin yhdisteihin tai molekyylien telakointia, kun proteiinikohteen rakennetiedot tunnetaan.

Virtuaaliseulontamenetelmät luokitellaan usein joko ligandipohjaisiin (engl. *Ligand-Based Virtual Screening*, LBVS) tai proteiinin rakenteeseen pohjautuviin (engl. *Structure-Based Virtual Screening*, SBVS) menetelmiin. LBVS-menetelmissä tarkastellaan vain kemiallisia yhdisteitä (ligandeja) eikä proteiinikohteita. SBVS-menetelmissä tarkastellaan, kuinka yhdiste vuorovaikuttaa proteiinin rakenteen kanssa.

Kun QSAR-menetelmiä käytetään virtuaaliseulontaan, on yleistä käyttää luokitusmalleja ("aktiivinen" tai "inaktiivinen")

[92] http://www.r-project.org

[93] http://www.cs.waikato.ac.nz/ml/weka

[94] https://www.3ds.com/products-services/biovia/products/data-science/pipeline-pilot

[95] https://www.knime.com

sekä kvantitatiivisia ennustusmalleja, joilla yritetään ennustaa sitoutumisen voimakkuutta.

Ennustavien mallien arviointi

Ennustavan mallin tehokkuuden arviointi voi olla monimutkaista useista syistä:

- Monet koulutukseen ja validointiin käytettävät tietojoukot ovat vinossa. Toisin sanoen niissä paljon enemmän inaktiivisia kuin aktiivisia yhdisteitä. Aktiivisten ja inaktiivisten suhde voi olla jopa 1:1000. Tällaisten sarjojen tilastollinen käsittely voi olla ongelmallista, koska jos malli ennustaa joka kerta inaktiivisen yhdisteen, on tulos oikea 99,9 %). Tätä voidaan korjata koulutusvaiheessa ottamalla näytteitä inaktiivisista yhdisteistä, ja vastaavasti validointivaiheessa luomalla täydellinen sekaannusmatriisi (lue alta ja vertaa kuvaan 14).

- Ristiinvalidointi osoittaa, kuinka hyvin malli ennustaa yhdisteitä samasta tietojoukosta, joita käytettiin mallin kouluttamiseen. Se ei kerro, miten hyvin malli toimii ennakoivasti täysin tuntemattomille yhdisteille. Tämä johtuu siitä, että mallin kouluttamiseen ja validointiin käytetyillä yhdisteillä on tietty soveltamisala (tai peitto deskriptoriavaruudessa). Tämä ei välttämättä ole sama kuin yhdisteillä, joihin mallia tullaan käyttämään. Siksi on tärkeää pystyä määrittämään, onko testattava yhdiste yhteensopiva mallin soveltamisalan kanssa. Tämä voidaan tehdä esimerkiksi vertaamalla sen samankaltaisuutta koulutuksessa ja validoinnissa käytettyihin yhdisteisiin.

- Mallin kouluttamiseen käytettyjen aktiivisuusarvojen määrittämiskokeiden virhesuhteet voivat olla tuntemattomia.

QSAR-mallien arvioimiseksi kehitettyjä hyviä käytänteitä:

- Sekaannusmatriisi luodaan mallin validointivaiheessa. Sekaannusmatriisi osoittaa, miten aktiivinen/inaktiivinen todellinen data vertautuu ennustettuun aktiiviseen/inaktiiviseen dataan. Tarkemmin sanottuna, se osoittaa ennustettujen aktiivisten ja aidosti aktiivisten yhdisteiden määrät (todelliset positiiviset, TP), ennustettujen inaktiivisten, jotka ovat todellisuudessa aktiivisia (väärät negatiiviset, VN), ennustettujen aktiivisten, jotka ovat todellisuudessa inaktiivisia (väärät positiiviset, VP) ja ennustetut inaktiiviset, jotka ovat myös todellisuudessa inaktiivisia (todelliset negatiiviset, TN) (ks. alta esimerkkitaulukko). Tavoitteena on maksimoida TP- ja TN-määrät ja minimoida VP- ja VN-määrät. Huomaa, että vinossa tietojoukossa (kuten yllä oleva suhde 1:1000), jonka malli ennusti aina inaktiivisen, olisi matriisissa TP-nopeus nolla ja VP-nopeus korkea. Tämä toisi mallin haasteen selkeästi esille.

	Ennustetut aktiiviset	**Ennustetut inaktiiviset**
Aidosti aktiiviset	Todelliset positiiviset	Väärät negatiiviset
Aidosti inaktiiviset	Väärät positiiviset	Todelliset negatiiviset

- Sekaannusmatriisilla voidaan tehdä useita hyödyllisiä tilastollisia analyysejä, kuten esimerkiksi tarkkuus (engl. *precision*) (TP/TP+VP; aktiivisten yhdisteiden osuus aktiivisiksi ennustetuista ja havaituista), palautus (engl. *recall*) (TP/TP+VN; aidosti aktiiviseksi havaittujen yhdisteiden osuus kaikista aktiivista) ja f-pisteet (tarkkuuden ja palautuksen harmoninen keskiarvo).

- Erityisesti virtuaaliseulontasovelluksissa, joissa malli palauttaa paremmuusjärjestykseen määritetyn luettelon, on hyödyllistä piirtää ROC-käyrä (engl. *Receiver Operating Curve*). ROC-käyrä osoittaa todellisten ja väärien positiivisten määrät luettelon laskiessa. Se on siis tarkkuuden ja palautuksen graa-

finen versio. Tämä kuvaa, kuinka hyvin malli palauttaa aktiivisia yhdisteitä lähellä listan kärkeä verrattuna vääriin positiivisiin tuloksiin. Numeerinen mitta voidaan laskea myös käyrän alla olevana pinta-alana (engl. *Area Under the Curve* (AUC)). ROC-käyrien soveltamisesta keminformatiikassa voi lukea lisää artikkelista Jain & Nicholls (2008)[96].

- On pyrittävä ymmärtämään mallien rajauksia ja soveltamisalueita. Lisäksi niiden koulutus- ja validointisarjat tulisi julkaista, jotta niiden soveltuvuutta tiettyihin yhdisteisiin voitaisiin arvioida.

Tehtävät

34. Millaisten tekijöiden perusteella voisit päättää, mitä epälineaarista menetelmää tulisi käyttää tiettyyn ongelmaan?
35. Voidaanko mallin rakentamiseen käyttää sekä binäärisiä että muunlaisia deskriptoreita?
36. Miten mallin rajoituksia voidaan analysoida?
37. Katso alla olevasta sekaannusmatriisista validointisarja, joka sisältää 1 000 yhdistettä. Yhdisteistä 30 on aidosti aktiivisia. Onko malli mielestäsi hyvä?

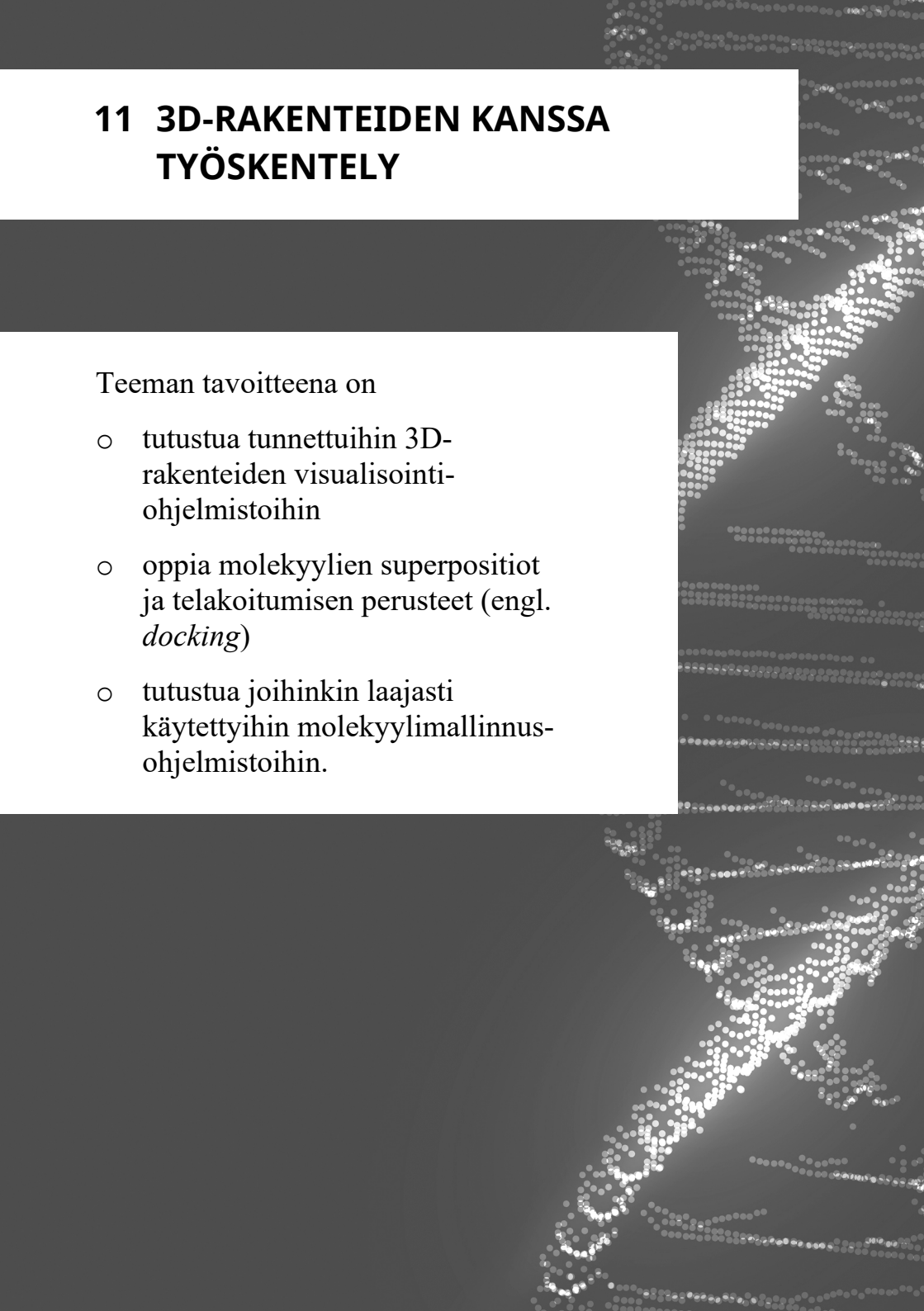
	Ennustetut aktiiviset	**Ennustetut inaktiiviset**
Aidosti aktiiviset	14	16
Aidosti inaktiiviset	2	968

[96] https://doi.org/10.1007/s10822-008-9196-5

11 3D-RAKENTEIDEN KANSSA TYÖSKENTELY

Teeman tavoitteena on

- tutustua tunnettuihin 3D-rakenteiden visualisointi-ohjelmistoihin
- oppia molekyylien superpositiot ja telakoitumisen perusteet (engl. *docking*)
- tutustua joihinkin laajasti käytettyihin molekyylimallinnus-ohjelmistoihin.



Johdanto

Teemassa 6 käsiteltiin 3D-rakenteiden esittämistä ja karakterisointia, mukaan lukien konformeerien tuottaminen, energian minimointi sekä 3D-farmakoforien ja samankaltaisuuden hakeminen. Tässä teemassa tarkastellaan lyhyesti molekyylien visualisointia ja sitä, kuinka 3D-rakenteita voidaan käyttää erilaisissa laskennallisissa sovelluksissa joko erikseen tai yhdessä proteiinikohderakenteen kanssa.

3D-rakenteiden ja proteiinien visualisointi

Yleisimmät ja yksinkertaisimmat tavat visualisoida 3D-rakenteita ovat kalottimallit ja pallotikkumallit. Mielenkiintoista on, että molekyylivisualisoinnit kehitettiin hyvin varhaisessa vaiheessa (1960- ja 70-luvuilla), ja kun korkealaatuinen grafiikka yleistyi 1990-luvulla, oli 3D-visualisoiminen yksi huippuluokan grafiikkatyöasemien tärkeimmistä sovelluksista. Yhdisteiden ja erityisesti proteiinikohteisiin sitoutuneiden yhdisteiden visualisointi (esim. röntgenkristallografiassa) voi olla erittäin hyödyllistä antaen muun muassa tietoa siitä, mistä yhdistettä voidaan muokata saaden se sitoutumaan voimakkaammin kohteeseen.

Nykyisin on tarjolla useita hyviä ohjelmistoja, joilla molekyylejä voi visualisoida, pyörittää ja käännellä. Monet näistä ovat ilmaisia. Osaa niistä ajetaan verkkoselaimella ja toiset asennetaan tietokoneelle. Lähes kaikilla voi tuottaa kuvia. Suurimmalla osalla voi työskennellä sekä proteiinien että pienten molekyylien kanssa. Muutamia merkittäviä ovat:

- JMOL[97]
- Rasmol[98]
- PyMOL[99]

[97] http://jmol.sourceforge.net

[98] http://www.openrasmol.org

- Cn3D[100]
- Mollycule[101].

Verkosta löytyy useita tutoriaaleja, erityisesti Jmolille:

- UWEC:n tutoriaali Jmolin peruskäyttöön[102]
- Wileyn johdatus Jmolin käyttöön[103]
- California Lutheranin johdatus Jmol-skiptaamiseen[104].

Molekyylien superpositio

Molekyylien superpositio tarkoittaa kahden tai useamman molekyylin kohdistamista 3D:ssä siten, että ne peittävät optimaalisesti jotakin. Ne kohdistetaan joko toistensa tai jäykän vertailumolekyylin kanssa. Kohdistus tapahtuu kiertämällä ja kääntämällä molekyylejä sekä taivuttamalla ja kiertämällä sidoksia erilaisten konformeerien luomiseksi. Superpositiointi voi olla edellytys 3D-samankaltaisuuden etsinnälle, farmakoforien havaitsemiselle ja 3D QSAR:lle.

Superpositiointi on optimoinnin sovellus, ja se voi voidaan toteuttaa monella eri menetelmällä, kuten esimerkiksi jyrkimmän suunnan -menetelmällä, geneettisillä algoritmeilla, simuloidulla hehkutuksella tai Monte Carlo -simulaatiolla. Myös muita menetelmiä voidaan soveltaa, kuten esimerkiksi muotoon pohjautuvaa päällekkäisyyttä (engl. *shape-based overlay*) (ks. esim. OpenEye ROCS).[105]

---

[99] https://pymol.org

[100] https://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml

[101] http://mollycule.graemebell.net

[102] https://www.chem.uwec.edu/JmolTut

[103] https://www.wiley.com/legacy/college/boyer/0471661791/structure/jmol_intro/jmol_intro.htm

[104] http://earth.callutheran.edu/Academic_Programs/Departments/BioDev/omm/scripting/molmast.htm

[105] http://eyesopen.com/rocs

3D QSAR

3D QSAR voidaan tehdä yksinkertaisesti käyttämällä 3D-deskriptoreita 2D:n sijaan. Mutta on myös olemassa muita tapoja tehdä QSAR 3D-muodossa. Yksi hyvin tunnettu menetelmä on CoMFA (engl. *Comparative Molecular Field Analysis*). CoMFA optimoi useiden rakenteiden päällekkäisyyden kenttien avulla sitoutumiseen liittyvien ominaisuuksien (esim. sähköstaattisuus, steerisyys, hydrofobisuus) perusteella. Se löytää päällekkäisistä kentistä yhtäläisyyksiä ja korreloi näitä alueita aktiivisuuteen. CoMFA:n käyttö edellyttää, että rakenteet ovat jo kohdistettu. Tuloksia voidaan käyttää ennakointiin tai visualisointiin. Lisätietoja löydät CoMFA-oppaasta.[106]

Molekyylien telakointi

Molekyylitelakoinnin tavoitteena on ennustaa, kuinka yhdisteet (tässä kontekstissa ligandit) voivat sitoutua proteiinien aktiivisiin kohtiin. Yleensä tarkoituksena on arvioida niiden proteiinin toiminnan estämispotentiaalia. Useimmat telakointiohjelmistot ottavat syötteenä ligandin 3D-rakennetiedoston ja proteiinikohteen 3D-rakenteen, jossa aktiivinen kohta on tunnistettu. Sitten ne kiertävät ja kääntävät aktiivisessa kohteessa olevia yhdisteitä taivuttamalla pyöriviä sidoksia, kunnes jokin sitoutumiseen liittyvä ominaisuus on minimoitu. Tämä pisteytystoiminto koostuu usein esimerkiksi vetysitoutumisen, sähköstaattisesta vuorovaikutuksen, muodon ja hydrofobisen vuorovaikutuksen yhdistelmästä. Pisteytystoiminnon lopullista arvoa pidetään yleensä telakointimenestyksen mittarina. Vaikka telakointialgoritmit ovat osoittautuneet varsin hyviksi aktiivisten kohtien yhdisteiden sitoutumissuuntauksien toistamisessa (esim. testit, joissa käytetään RMSD-menetelmää sitoutuneista ligandeista kiderakenteissa), pisteytystoiminnon lopullisen arvon suhde sitoutumisaffiniteettiin ei ole

[106] http://life.nthu.edu.tw/~lslth/QSARandCoMFA/sybyl/sybyl/qsar/qsar_tut26.html

niin selkeä. Saatavilla on sekä ilmaisia että kaupallisia telakointityökaluja.

Ilmaisia ohjelmistoja:

- UCSF Dock[107]
- Scripps Autodock[108]
- DockingServer[109] (verkkopalvelu, ilmaiset rajoitetut palvelut).

Kaupallisia ohjelmistoja:

- Schroedinger Glide[110]
- CCDC GOLD[111]
- SeeSAR[112]
- OpenEye OEDocking[113].

Molekyylitelakointia voidaan käyttää kahteen asiaan. Sen avulla voi ennustaa yhdisteiden kohdistumista proteiinikohteeseen (esim. myöhempää visualisointia varten) tai suuren mittakaavan virtuaaliseulontaan. Virtuaaliseulonnassa verrataan suurta määrää yhdisteitä proteiinin kohteeseen todellisen seulontakokeen sijaan.

Viimeaikaisessa tutkimuksessa on pyritty vertailemaan näitä telakointimenetelmiä arvioimalla telakointiasentojen tarkkuutta suhteessa kiderakenteisiin (käyttäen RMS-poikkeamaa) sekä selvittämällä, miten hyvin pisteytystoiminto korreloi sitoutumisaffiniteettiin (mitattuna ROC-käyrillä). Esimerkiksi Journal of Chemical Information and Modeling -lehden artikkeli Cross et al.

[107] http://dock.compbio.ucsf.edu

[108] http://autodock.scripps.edu

[109] https://www.dockingserver.com/web

[110] https://www.schrodinger.com/glide

[111] https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold

[112] https://www.biosolveit.de/products/seesar

[113] https://www.eyesopen.com/oedocking

(2009)[114] osoittaa, että testatuissa aineistoissa Glide ja Surflex ylittävät muut menetelmät molemmissa laskelmissa.

Molekyylimallinnusohjelmistoja

Saatavilla on useita molekyylimallinnustyökaluja, joilla voidaan tehdä superpositiointia, 3D QSAR:a, telakointia ja kvanttilaskelmia. Tässä muutama esimerkki:

- BIOVIA Discovery Studio Visualizer[115]
- Chimera[116]
- Ghemical[117]
- CCP4[118]
- PCModel[119].

[114] https://doi.org/10.1021/ci900056c

[115] https://discover.3ds.com/discovery-studio-visualizer-download

[116] https://www.cgl.ucsf.edu/chimera

[117] http://www.bioinformatics.org/ghemical/ghemical/index.html

[118] https://www.ccp4.ac.uk

[119] http://www.serenasoft.com

Tehtävät

38. Käytä RSCB proteiini tietopankin JSmol-näkymää visualisoidaksesi HIV-proteaasin estäjälääkkeen kiderakennetta, joka on sitoutunut HIV-proteaasiproteiiniin. Aloita tältä sivulta.[120] Vertaa Cartoon-visualisointia Ligands and Pocket -visualisointiin (valitaan Style-valikosta). Kuinka näitä visualisointeja voidaan käyttää eri tarkoituksiin?
39. Tutustu Jmoliin Wileyn Introduction to JMOL -sivuston[121] avulla.[122]
40. Kuinka voisit arvioida eri telakointimenetelmien soveltuvuutta tietylle tietojoukolle ja proteiinikohteelle?

[120] http://www.rcsb.org/pdb/explore/jmol.do?structureId=1HVI&bionumber=1

[121] http://www.wiley.com/legacy/college/boyer/0471661791/structure/jmol_intro/jmol_intro.htm

[122] Suomentajan kommentti: Wileyn sivustoa ei ole päivitetty Jmol-pohjainen ja tarvitsee sen vuoksi Java-tuen. Vaihtoehtoisesti voit tutustua teknologiaan javattoman JSmol-pohjaisen Jmol-tutoriaalin avulla: https://www.andrew.cmu.edu/user/rule/jsmol/jsmol_tutorial.html

12 KEMINFORMATIIKAN OHJELMISTOKEHITYS

Teemassa tarkastellaan keminformatiikan ohjelmistokehitystä. Tavoitteena on tutustua soveltuviin ohjelmointityökaluihin ja työnkulkuohjelmistoihin.

Keminformatiikan ohjelmointityökalut

Onneksi on saatavilla useita ilmaisia ja avoimen lähdekoodin työkaluja, jotka sisältävät kirjastoja yleisimmille keminformatiikkatoiminnoille, kuten rakenteiden esittämiselle ja hakemiselle. Näin on mahdollista kehittää nopeasti uusia ohjelmistoja ilman, että ”pyörää tarvitsee keksiä uudestaan".

Chemistry Development Kit (CDK)[123] on keminformatiikassa laajasti käytetty avoimen lähdekoodin Java-työkalupaketti. CDK:n verkkosivuston mukaan sillä on yli 50 kehittäjää maailmanlaajuisesti. CDK tarjoaa laajan valikoiman ominaisuuksia, mukaan lukien 2D-rakenteiden syöttäminen, esittäminen ja kuvailu, tiedostojen ja lineaarinotaatioiden konvertointi, 3D-renderointi, yksinkertaisiin kuvaajiin perustuva virtuaaliseulonta ja käyttöliittymän R-ohjelmistoon, 3D-mallien rakentaminen, alirakenteen haku, NMR-ennustus ja rakenteiden generointi. Ohjelmistossa on myös käyttöliittymä BioJava bioinformatics -työkalupakettiin.

CDK on esitelty artikkelissa Steinbeck et al. (2003)[124], joka julkaistiin Journal of Chemical Information and Computer Sciences -lehdessä, ja päivitykset Current Pharmaceutical Design -lehden artikkelissa Steinbeck et al. (2006)[125].

OpenBabel[126] on valmiiden ohjelmistojen "työkalupakki" ja keminformatiikan C/C++-työkalu, jossa on kääreitä (engl. *wrapper*) muille kielille. OpenBabel perustuu kaupallisen, mutta akateemikoille ilmaisen OEChemin[127] aikaisempaan versioon. OpenBabel mahdollistaa rakennetiedostojen konvertoinnin, konformeerien generoinnin, energialaskennan, minimoinnin ja monia muita toimintoja. OpenBabelissa on myös monia keminformatiikan 2D-

[123] http://cdk.sourceforge.net

[124] https://doi.org/10.1021/ci025584y

[125] https://doi.org/10.2174/138161206777585274

[126] http://openbabel.org

[127] http://www.eyesopen.com/oechem-tk

ja 3D-toimintoja. Ohjelmistokokonaisuus on esitelty Journal of Cheminformatics -lehdessä vuonna 2011. (O'Boyle ym., 2011)[128]

Chemistry Descriptors Library (CDL)[129] on C ++-kirjasto, joka tarjoaa laajan valikoiman keminformatiikan toimintoja, mukaan lukien rakenteiden esittäminen ja konvertointi, deskriptorien ja sormenjälkien luominen, pKa-ennustaminen ja synteettisen saavutettavuuden arviointi. Kirjasto on esitelty vuoden 2008 Journal of Chemical Information and Modeling -lehden artikkelissa Sykora & Leahy (2008)[130].

Lisäksi on olemassa useita muita työkaluja, kuten .NET-keminformatiikkatyökalu MolEngine[131], koneoppimisen ja keminformatiikan työkalu RDKit[132], avoimen lähdekoodin C++-kirjasto ChemKit[133] molekyylimallinnukseen, visualisointiin ja keminformatiikkaan, Indigo Toolkit[134] on avoimen lähdekoodin C++-keminformatiikkatyökalu ja Maya ChemTools[135] (joukko keminformatiikkaan soveltuvia Perl-skriptejä).

Työnkulkuohjelmistot

Työnkulkuohjelmistot sallivat mukautettujen prosessien ja "ohjelmistojen" luomisen hyödyntämällä työnkulkuja tai olemassa olevia ohjelmia. Tämä mahdollistaa mukautettujen tehtävien suorittamisen ilman ohjelmointikokemusta.

Tällaisia työkaluja käytetään laajalti uusien lääkeaineiden löytämiseen liittyvässä laskennallisessa tutkimuksessa. Suosituin

[128] https://doi.org/10.1186/1758-2946-3-33

[129] http://cdelib.sourceforge.net/doc/index.html

[130] https://doi.org/10.1021/ci800135h

[131] https://www.scilligence.com/web/molengine

[132] http://rdkit.org

[133] http://wiki.chemkit.org

[134] https://lifescience.opensource.epam.com/indigo

[135] http://www.mayachemtools.org

kaupallinen ohjelmisto SciTegicin Pipeline Pilot[136], johon on integroitu useita keminformatiikan, bioinformatiikan ja yleisen tietojenkäsittelyn moduuleja. Muita ilmaisia ja/tai avoimen lähdekoodin keminformatiikkaan soveltuvia työnkulkuohjelmistoja ovat Knime[137], CDK Taverna (Truszkowski ym., 2011)[138] ja AZorange (Stålring ym., 2011)[139], joka on Orange-tiedonlouhinnan ja graafisen ohjelmoinnin ympäristön lisäosa.

Lisäksi saatavilla on monia muita työkaluja, kuten esimerkiksi R-tilasto-ohjelmisto. R-ohjelmistoon saa keminformatiikan toimintoja RCDK-rajapinnan avulla.[140]

Tehtävät

41. Millä perusteilla tekisit päätöksen siitä, mitä ohjelmointityökaluja käytetään?
42. Mitkä ovat työnkulkuohjelmistojen käytön edut ja haitat?

[136] https://en.wikipedia.org/wiki/Pipeline_Pilot

[137] http://www.knime.org

[138] https://doi.org/10.1186/1758-2946-3-54

[139] https://doi.org/10.1186/1758-2946-3-28

[140] http://cran.r-project.org/web/packages/rcdk/index.html

13 SEURAAVAKSI: MOOC-KURSSEJA JA MUITA VERKKO-OPPIMATERIAALEJA

Oppaan viimeisessä teemassa tutustutaan lyhyesti resursseihin, joiden avulla lukija voi laajentaa ja syventää keminformatiikan osaamistaan.

Huomaa, että tämä opas on keskittynyt keminformatiikan "ytimeen", mutta tulevana keminformatiikan ammattilaisena sinun on osattava tilastotieteitä, tiedon visualisointia ja ymmärtää lääkkeiden kehittämisprosessia.

Teema sisältää suosituksia MOOC-kursseista ja verkkomateriaaleista edellä mainitun osaamisen kehittämiseksi.

Keminformatiikan "ydin"

Henry Stewartin Introduction to Cheminformatics -sivulla on videoita keminformatiikan ydinalueista, kuten molekyylimallinnuksesta, algoritmeista, tietokannoista ja koneoppimisesta.[141]

Laajemmalla kemiallisen tiedonkäsittelyn aiheesta on tarjolla verkkoresursseja, kuten Chemical Information Sources Wiki[142].

MOOC-kursseja

Keminformatiikan opiskeluun on tarjolla MOOC-kursseja. Kurssit tarjoavat yleiskatsauksia keminformatiikan eri aiheista. Esimerkiksi Coursera[143] on tarjonnut kursseja:

- Computing for Data Analysis
- Data Analysis
- Network Analysis in Systems Biology
- Drug Discovery
- Intermediate Organic Chemistry
- Introduction to Systems Biology
- Web Intelligence and Big Data
- Bioinformatics Algorithms Part I.

Muita huomioitavia MOOC-kursseja ovat:

- Introduction to Biology – The Secret of Life (edX[144])
- Semantic Web Techologies (OpenHPI[145])
- Indianan yliopiston Information Visualization MOOC[146].

---

[141] https://hstalks.com/playlist/582/introduction-to-cheminformatics

[142] http://en.wikibooks.org/wiki/Chemical_Information_Sources

[143] https://www.coursera.org

[144] https://www.edx.org

[145] http://openhpi.org

[146] http://ivmooc.cns.iu.edu

LÄHTEET

Apodaca, R. (2010, syyskuuta 28). *A Brief Introduction to Lawson Numbers*. Depth-First. http://depth-first.com/articles/2010/09/28/a-brief-introduction-to-lawson-numbers

Ashton, M. J., Jaye, M. C., & Mason, J. S. (1996). New perspectives in lead generation II: Evaluating molecular diversity. *Drug Discovery Today*, *1*(2), 71–78. https://doi.org/10.1016/1359-6446(96)89091-X

Banville, D. L. (Toim.). (2008). *Chemical Information Mining: Facilitating Literature-Based Discovery* (0 p.). CRC Press. https://doi.org/10.1201/9781420076509

Barnard, J. M., & Downs, G. M. (1992). Clustering of chemical structures on the basis of two-dimensional similarity measures. *Journal of Chemical Information and Computer Sciences*, *32*(6), 644–649. https://doi.org/10.1021/ci00010a010

Bayada, D. M., Hamersma, H., & van Geerestein, V. J. (1999). Molecular Diversity and Representativity in Chemical Databases. *Journal of Chemical Information and Computer Sciences*, *39*(1), 1–10. https://doi.org/10.1021/ci980109e

Brown, F. (1998). Chemoinformatics: What is it and how does it impact drug discovery. *Annual Reports in Medicinal Chemistry*, *33*, 375–384.

Brown, R. D. (1996). Descriptors for diversity analysis. *Perspectives in Drug Discovery and Design*, *7*(1), 31–49. https://doi.org/10.1007/BF03380180

Chen, W. L. (2006). Chemoinformatics: Past, Present, and Future†. *Journal of Chemical Information and Modeling*, *46*(6), 2230–2255. https://doi.org/10.1021/ci060016u

Cross, J. B., Thompson, D. C., Rai, B. K., Baber, J. C., Fan, K. Y., Hu, Y., & Humblet, C. (2009). Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *Journal of Chemical Information and Modeling*, *49*(6), 1455–1474. https://doi.org/10.1021/ci900056c

Dalke, A. (2003, lokakuuta 15). *WLN -- History of Chemical Nomenclature*. Dalke scientific – More Science. Less Time. http://www.dalkescientific.com/writings/diary/archive/2003/10/15/WLN.html

Downs, G. M., & Barnard, J. M. (2003). Clustering Methods and Their Uses in Computational Chemistry. Teoksessa K. Lipkowitz & D. Boyd (Toim.), *Reviews in Computational Chemistry* (ss. 1–40). John Wiley & Sons, Ltd. https://doi.org/10.1002/0471433519.ch1

Dunbar, J. B. (1996). Cluster-based selection. *Perspectives in Drug Discovery and Design*, *7*(1), 51–63. https://doi.org/10.1007/BF03380181

Evans, M. (2011, elokuuta 31). Chemoinformatics Curiosities: The Morgan Algorithm. *Cheersical Education*.

https://graphiteworks.wordpress.com/2011/08/31/chemoinformatics-curiosities-i-the-morgan-algorithm

Gasteiger, J., & Engel, T. (Toim.). (2003). *Chemoinformatics: A textbook*. Wiley-VCH.

Jain, A. N., & Nicholls, A. (2008). Recommendations for evaluation of computational methods. *Journal of Computer-Aided Molecular Design*, *22*(3), 133–139. https://doi.org/10.1007/s10822-008-9196-5

Jiao, D., & Wild, D. J. (2009). Extraction of CYP Chemical Interactions from Biomedical Literature Using Natural Language Processing Methods. *Journal of Chemical Information and Modeling*, *49*(2), 263–269. https://doi.org/10.1021/ci800332w

Leach, A. R., & Gillet, V. J. (2007). *An introduction to chemoinformatics*. Springer. http://dx.doi.org/10.1007/978-1-4020-6291-9

Lin, J. (2009). Is searching full text more effective than searching abstracts? *BMC Bioinformatics*, *10*(1), 46. https://doi.org/10.1186/1471-2105-10-46

Lynch, M. F., & Holliday, J. D. (1996). The Sheffield Generic Structures Projecta Retrospective Review. *Journal of Chemical Information and Computer Sciences*, *36*(5), 930–936. https://doi.org/10.1021/ci9501731

Martin, Y. C. (1996). Challenges and prospects for computational aids to molecular diversity. *Perspectives in Drug Discovery and Design*, *7*(1), 159–172. https://doi.org/10.1007/BF03380186

McIntosh, T., & Curran, J. R. (2009). Challenges for automatically extracting molecular interactions from full-text articles. *BMC Bioinformatics*, *10*(1), 311. https://doi.org/10.1186/1471-2105-10-311

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, *3*(1), 33. https://doi.org/10.1186/1758-2946-3-33

O'Donnell, T. J. (2009). *Design and Use of Relational Databases in Chemistry*. CRC Press.

Oprea, T. I., Tropsha, A., Faulon, J.-L., & Rintoul, M. D. (2007). Systems chemical biology. *Nature Chemical Biology*, *3*(8), 447–450. https://doi.org/10.1038/nchembio0807-447

Shemetulskis, N. E., Dunbar, J. B., Dunbar, B. W., Moreland, D. W., & Humblet, C. (1995). Enhancing the diversity of a corporate database using chemical database clustering and analysis. *Journal of Computer-Aided Molecular Design*, *9*(5), 407–416. https://doi.org/10.1007/BF00123998

Southan, C., Várkonyi, P., & Muresan, S. (2009). Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *Journal of Cheminformatics*, *1*(1), 10. https://doi.org/10.1186/1758-2946-1-10

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *Journal of Chemical Information and Computer Sciences*, *43*(2), 493–500. https://doi.org/10.1021/ci025584y

Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., & Willighagen, E. L. (2006). Recent Developments of the Chemistry Development Kit (CDK)—An Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design*, *12*(17), 2111–2120. https://doi.org/10.2174/138161206777585274

Stewart, K. D., Shiroda, M., & James, C. A. (2006). Drug Guru: A computer software program for drug design using medicinal chemistry rules. *Bioorganic & Medicinal Chemistry*, *14*(20), 7011–7022. https://doi.org/10.1016/j.bmc.2006.06.024

Stålring, J. C., Carlsson, L. A., Almeida, P., & Boyer, S. (2011). AZOrange—High performance open source machine learning for QSAR modeling in a graphical programming environment. *Journal of Cheminformatics*, *3*(1), 28. https://doi.org/10.1186/1758-2946-3-28

Sykora, V. J., & Leahy, D. E. (2008). Chemical Descriptors Library (CDL): A Generic, Open Source Software Library for Chemical Informatics. *Journal of Chemical Information and Modeling*, *48*(10), 1931–1942. https://doi.org/10.1021/ci800135h

Truszkowski, A., Jayaseelan, K. V., Neumann, S., Willighagen, E. L., Zielesny, A., & Steinbeck, C. (2011). New developments on the cheminformatics open workflow environment CDK-Taverna. *Journal of Cheminformatics*, *3*(1), 54. https://doi.org/10.1186/1758-2946-3-54

Turner, D. B., Tyrrell, S. M., & Willett, P. (1997). Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *Journal of Chemical Information and Computer Sciences*, *37*(1), 18–22. https://doi.org/10.1021/ci960463h

Wild, D. J., & Blankley, C. J. (2000). Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering. *Journal of Chemical Information and Computer Sciences*, *40*(1), 155–162. https://doi.org/10.1021/ci990086j

Wild, David J. (2009). Grand challenges for cheminformatics. *Journal of Cheminformatics*, *1*(1), 1. https://doi.org/10.1186/1758-2946-1-1

Willett, P. (2011). Chemoinformatics: A history. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *1*(1), 46–56. https://doi.org/10.1002/wcms.1

Willett, P., Barnard, J. M., & Downs, G. M. (1998). Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences*, *38*(6), 983–996. https://doi.org/10.1021/ci9800211

LIITE 1. TEHTÄVIEN VASTAUKSET

Suurin osa tehtävistä on avoimia, joten liitteessä esitetyt vastaukset edustavat vain yhtä esimerkkiä monista pätevistä vastauksista.

Teema 1: Keminformatiikan historia ja nykytila

1. Selkeä ero olisi se, että bioinformatiikka keskittyy proteiineihin ja suurempiin kokonaisuuksiin, kun taas keminformatiikka keskittyy proteiineihin ja pienempiin yksiköihin. Kun nämä kaksi kenttää kehittyvät, tämä jaottelu hämärtyy. Yksi keskeinen ero on, että bioinformatiikassa kemiallisia yhdisteitä tai lääkkeitä käsitellään usein "päätepisteinä", kun taas keminformatiikassa niitä pidetään portteina ominaisuuksien, funktionaalisten ryhmien, atomien, sidosten, ominaisuuksien jne. maailmaan. Kulttuurisesti nämä kaksi alaa ovat kehittyneet melko eri tavalla. Bioinformatiikka on kehittynyt akateemisen yhteisön kautta, ja suurin osa tutkimuksesta tehdään julkisella sektorilla. Keminformatiikan kehitys on ollut pääosin teollisuuden ohjaamaa, mutta viime vuosikymmenen aikana akateeminen läsnäolo on ollut paljon suurempaa.
2. Neljä aluetta ovat: lääkeainetutkimuksen nopeuttaminen, vihreä kemia ja ilmaston lämpeneminen, elämän ymmärtäminen kemiallisesta näkökulmasta ja maailman kemiallisten ja biologisten tietojen verkoston saatavuuden ja tulkinnan mahdollistaminen.
3. Todellinen haaste on kemiallisen ja keminformaattisen datan integrointi muiden tieteenalojen (esim. biologian, bioinformatiikan, kokeellisen ja simuloidun) datoihin. Näiden alojen rajapinnoissa tarvitaan uutta tutkimusta.
4. Keminformatiikkaa on sovellettu muun muassa materiaalitieteessä, energiateollisuudessa, maatalouskemikaaleissa ja kemikaalien toimituksessa.
5. Tähän sinun tulee vastata itse.

Teema 2: Kemian 2D-rakenteiden esittäminen tietokoneella

6. PubChemistä haettuna SMILES: CC(C)CC1=CC=C(C=C1)C(C)C(=O)O ja InChI: InChI=1S/C13H18O2/c1-9(2)8-11-4-6-12(7-5-11)10(3)13(14)15/h4-7,9-10H,8H2,1-3H3,(H,14,15). Huomaa, että tälle yhdisteelle on monta mahdollista SMILES-kaavaa.

7. Ensimmäinen yhdiste on L-Dopa, Parkinsonin taudin lääke. Toinen on stereoisomeeri, D-Dopa. Ensimmäisen osan verkkohaun pitäisi löytää molemmat, mutta koska L-Dopa on tunnetuin ja viitatuin, se hallitsee hakutuloksia.

8. Yhdiste on bentsimidatsoli. Daylight-ohjelmisto aromaattistaa yhdisteet automaattisesti käyttämällä Hückelin 4n+2-sääntöä riippuen siitä, miten ne syötettiin. Siksi se pitää molempia renkaita aromaattisina.

9. Ensimmäisen vaihtoehdon SMILES antaa yksinkertaisesti eksplisiittisen aromatisoitumisen ensimmäiselle renkaalle, mutta ohjelmisto ratkaisee sen samalla rakenteella. Toinen vaihtoehto ei ole pätevä, koska se tekisi molemmista typpiatomeista täysin aromatisoidut (ei vetyä kummassakaan typessä) ja rikkoisi siten 4n+2-sääntöä. Oikea merkintä olisi c1ccc2c(c1)nc[nH]2. Joten yleissääntö on, että jos käytät aromaattista input-muotoa, kaikkien renkaiden tulee noudattaa 4n+2-sääntöä (ainakin Daylight-ohjelmistossa).

Teema 3: 2D-rakenteiden karakterisointi deskriptoreilla ja sormenjäljillä

10. Jos mitkään sanakirjan fragmenteista eivät koodaa ominaisuuksia, jotka eroavat sarjan molekyylien välillä, niitä ei voida erottaa sormenjäljestä.

11. Jos niiden välisiä eroja ei ole koodattu fragmenteissa, joita käytetään sormenjäljen luomiseen (esimerkiksi stereoisomeerit).

12. Kyllä, vaikka muutama asia on pidettävä mielessä. Ensinnäkin kullekin sormenjäljen kuvaajalle annettaisiin oletuksena sama painoarvo, mutta on esimerkiksi vaikea sanoa esimerkiksi, että LogP olisi

painotettava samalla tavalla kuin tietyn rakenteellisen piirteen läsnäolo. Toiseksi, jos sormenjälkeä käytetään samankaltaisuuden laskemiseen, jokainen deskriptori olisi normalisoitava samalle alueelle. Lisäksi tässä tapauksessa olisi käytettävä ei-binääristä samankaltaisuuskerrointa.

13. Jos tarkistat Wikipedia-sivun, huomaat, että vaikka statiinien katsotaan kuuluvan samaan luokkaan, ovat ne rakenteellisesti melko erilaisia. Lisäksi ne sisältävät paljon hiiltä sisältäviä polkuja, joten hajautettujen sormenjälkien avulla samankaltaisuuslaskelmat voivat ylikuormittua. Koska PubChemissä on monia yhdisteitä, jotka muistuttavat enemmän kutakin statiinia kuin toisiaan, statiinit eivät nouse hakutulosten kärkeen.

Teema 4: 2D-rakenteiden etsiminen ja tallentaminen tietokantaan

14. Rakenteen haku, ja vain, jos samaa kanonisointialgoritmia käytetään sekä kyselyyn että SMILES-kaavaan. Tässä tapauksessa voit suorittaa yksinkertaisia tekstikyselyjä SMILES-kaavalla.

15. Hiili (järjestysluku 6) on sitoutunut hiileen, jossa on kolme sidosta ([#6][CX3]). Yksi sidoksista on kaksoissidos happeen, joten hiili on sp^2-hybridisoitunut ([CX3](=O)). Hiiliatomin kolmas sidos on yksinkertainen sidos toiseen hiiliatomiin ([#6]), joten SMARTS-rakenne edustaa ketonia [#6][CX3](=O)[#6].

16. Teoksen kirjoittamisen aikaan PubChemissä oli 30 217 998 yhdistettä[147] ja ChemSpiderissa yli 26 miljoonaa[148].

[147] Suomentamisen aikana (26.11.2020) yhdisteitä oli 111 458 063 (https://pubchemdocs.ncbi.nlm.nih.gov/statistics).

[148] https://www.chemspider.com/DataSources.aspx

Teema 5: Kemiallisten reaktioiden käsittely tietokoneella

17. Reaktioyhtälö kuvaa reaktion alku- ja loppuvaiheet eli lähtöaineiden muuttumisen tuotteiksi. Reaktiomekanismi esittää yksityiskohtaisesti, miten prosessi tapahtuu. Kemiallinen muutos kuvaa reaktiossa tapahtuvan funktionaalisen ryhmän tai muun alirakenteen muutoksen.
18. Kemiallista muutosta
19. Esimerkiksi CC(=O)C.O=CCC>>CC(=O)CC(O)CC. Huomaa, että esimerkki on sekä validi reaktio-SMILES että SMIRKS.
20. Reaktiotiedot ovat pääosin julkaistussa kirjallisuudessa, joten ne on eristettävä ja kuratoitava manuaalisesti. Tämän vuoksi reaktiotietokannat ovat yleensä kaupallisia.

Teema 6: 3D-rakenteiden esittäminen tietokoneella

21. Argumentteja useiden konformeerien tallentamiselle: Algoritmien tehokkuus (ei tarvitse huolehtia molekyylien taipumisesta), ja kyky rajata tarkasteltavat konformeerit alimman energian mukaan.

    Argumentteja taipuisuuden algoritmiselle käsittelylle: Varastoinnin tehokkuus (ei tarvitse tallentaa useita konformeereja). Lisäksi voidaan sallia koko konformaatioavaruuden laajuinen haku kullekin algoritmille.
22. Bromibentseeni (Bromobenzene) = 15.837003 ja klooribentseeni (Chlorobenzene) = 15.271241.
23. Kun etsitään mahdollisia uusia yhdisteitä sitoutumaan proteiinin aktiiviseen kohtaan rajoittamatta hakua tiettyyn rakenteelliseen sarjaan, ja kun aktiivisen kohdan rakenne on tiedossa.

Teema 7: Kemialliset rakenteet verkossa ja tieteellisissä julkaisuissa

24. Teoksen kirjoitushetkellä molempien hakujen ylin osuma oli ChemSpider-tietue, mutta eri tietue kullekin haulle. Systemaattiselle nimihaulle ID 9889970 ja InChI-avainhaulle ID 10607847. Mo-

lemmat tulokset ovat valideja. Huomaa, että nimihaku palauttaa paljon enemmän "vääriä positiivisia" kuin InChI-haku.

25. Oletettavasti parasetamoli ja asetaminofeeni ovat toistensa synonyymiluetteloissa. Jokaisella voi olla päällekkäinen, mutta myös erilainen synonyymiluettelo, mikä johtaa osumien eri määrään.

26. IC_{50}- ja rakennetiedot eivät välttämättä sijaitse tekstin samassa osassa. Esimerkiksi rakennetta voidaan kuvata seuraavasti: rakenne 1, esitetty kuvassa 2 (kuvassa on yhdisteen 2D-rakennekaava), ja muualla on esitetty rakenteen 1 IC_{50}-arvot.

Teema 8: Keminformatiikka kemian kirjastoissa

27. Selkeitä valintoja olisivat kaupalliset työkalut, jotka antavat pääsyn reaktiotietokantoihin Scifinder Scholar (CASREACT) ja Reaxys (Beilstein). Kyselyn määrittäminen riippuu sen ilmaisutavasta. Kyselyllä voidaan etsiä tekstitermejä, tekijöitä, alirakenteita, kemiallista samankaltaisuutta tai näistä muodostettua yhdistelmää. Suuria julkisia reaktiotietokantoja ei ole olemassa, mutta kaupallista hakua voidaan täydentää ilmaisilla työkaluilla reagenssien saatavuuden seuraamiseksi (esim. ChemSpider tai eMolecules) tai reagenssien tai tuotteiden kemiallisten ominaisuuksien tutkimiseksi (esim. PubChem tai Molinsipration).

28. Tähän tehtävään ei ole hyvää vastausta. Se edellyttäisi todennäköisesti syvällisempää analyysiä siitä, mistä arvot ovat peräisin, ja mikä arvoista on luotettavin. Ole tarkkana, kun tunnistat, mitkä arvot ovat todella kokeellisia ja mitkä ennustettuja.

29. Ennakoivat työkalut, kuten Molinspiration-sivusto, tarjoavat pääsyn LogP-ennusteisiin. Ennusteet voidaan laatia yhdisteille, joita ei ole tietojoukoissa. Kokeelliset arvot hyvin samankaltaisille yhdisteille voidaan tunnistaa myös esimerkiksi PubChemin samankaltaisuushaulla.

Teema 9: Kemiallisen tietoaineiston analysointi klusteroinnin ja monimuotoisuuden avulla

30. Suhteellinen menetelmä, koska se on johdettu tietojoukon yhdisteiden samankaltaisuuslaskelmista ilman ulkoista yhteyttä.

31. Yksi lähestymistapa olisi klusteroida tietojoukko ensin *n* klusteriksi käyttämällä epähierarkkista menetelmää. Lukumäärä *n* tulisi valita siten, että kaikki tietojoukot olisivat riittävän pieniä Wardin menetelmälle. Toinen tapa olisi käyttää hierarkkista jakomenetelmää, kuten Divisive *K*-Meansia.

32. Ongelmana on määritellä objektiivisesti, mitä tarkoitetaan ”kemiallisella sarjalla”. Kemistit eivät välttämättä ole yhtä mieltä siitä, miten yhdisteet järjestetään sarjoiksi. Tämän vuoksi ei voi olla yhtä ihanteellista klusterointimenetelmää.

33. Vastaus riippuu siitä, analysoidaanko kattavuutta (jolloin vasemmalla oleva on monimuotoisempi) vai suhteellista monimuotoisuutta (jolloin oikealla puolella oleva on monimuotoisempi).

Teema 10: Kemiallisten yhdisteiden biologisen aktiivisuuden ennustaminen

34. Käytettävissäsi olevat menetelmät (esim. Bayesian, jos käytetään Pipelinen pilotia): ongelman tyyppi (luokittelu vs. jatkuva), tulkittavuus (jos tarkoituksen on ymmärtää, miksi ennuste tehdään), tehokkuus (kuinka kauan kestää harjoittelu ja ennustaminen) ja viritysmenetelmien monimutkaisuus (esim. neuroverkot ja SVM:t voivat sovittaa liikaa, jos niitä ei parametrisoida oikein).

35. Kyllä, mutta on huolehdittava siitä, että arvot normalisoidaan oikein (kuten klustereissa).

36. Yleismallien haasteena on, että niitä voidaan soveltaa laajasti, mutta niiden ennustustarkkuus on heikko. Erikoistuneilla malleilla on suppeampi soveltamisala, mutta ne ennustavat hyvin. Yksi lähestymistapa on rakentaa useita erikoistuneita malleja ja valita sitten malli, joka soveltuu parhaiten tietylle yhdisteelle (esim. samankaltaisuus validointijoukon kanssa).

37. Malli ennustaa oikein melkein kaikki inaktiiviset yhdisteet, ja ennustaa oikein hieman alle puolet (47 %) aktiivisista. Mallin hyvyyden arviointi riippuu näkökulmasta ja sovelluskohteesta. Verrattuna satunnaiseen valintaan (joka tunnistaisi aktiivista 3 %), se näyttää melko hyvältä.

Teema 11: 3D-rakenteiden kanssa työskentely

38. Sarjakuvamalli (cartoon) soveltuu proteiinin mittakaavan ja ligandien sitoutumisen visualisoimiseen. Esim. tässä tapauksessa voidaan tunnistaa proteiinin olevan dimeeri. Sen "läpät" avautuvat päästäen ligandin (johon inhibiittori sitoutuu) sisään. Ligandi- ja taskukuva (engl. *The Ligand and Pocket view*) mahdollistaa sitoutumiskohdan yksityiskohtaisen tarkastelun, kuten ligandin ja aminohappojen funktionaalisten ryhmien väliset vuorovaikutukset.

39. Ei edellytetä vastausta.

40. Kaksi testaustapaa ovat, kuinka hyvin ohjelma asemoi proteiinikohteessa olevia ligandeja, ja kuinka hyvin telakointipisteet korreloivat aktiivisuuden kanssa. Ensimmäistä voidaan testata käyttämällä tunnettuja telakointiasentoja (esim. kiderakenteista), eristämällä ligandi, minimoimalla sen rakenne ja antaa telakointiohjelmiston yhdistää ligandi aktiiviseen kohtaan uudelleen. Menetelmän rajoite on, että aktiivinen kohta on joustava ja ligandi sopi siihen täydellisesti jo alun perin. Jälkimmäistä voidaan testata ennustemallin kaltaisesti käyttämällä ristiinvalidointistrategioita.

Teema 12: Keminformatiikan ohjelmistokehitys

41. Päätökseen tekoon vaikuttaa muun muassa: kustannukset, tehdäänkö tiedostoista ja lähdekoodista avoimesti saatavaa, mitä ohjelmointikieltä käytetään ja millaisia tukipalveluita tarjotaan.

    Edut: syvällisiä ohjelmointitaitoa ei vaadita, toiminnallisuuksia voidaan kehittää valmiilla rakennuspalikoilla ja työkalut mahdollistavat visuaalisen työjonon mallinnuksen. **Haitat**: kehitetyt ohjelmistot ovat riippuvaisia suoritettavasta työnkulkuohjelmistosta, työnkulun malli rajoittaa joustavuutta ja voi olla vähemmän tehokas, kuin alusta alkaen tehty tai työkalupakilla kirjoitettu koodi.

LIITE 2. SUOMENTAJAN JÄLKISANAT

Haluan kiittää professori David Wildia oppaan käännösoikeuden myöntämisestä. Käännös on tärkeä keminformatiikan avaus suomalaiselle kemian opetukselle, jossa aihe on ollut tähän mennessä lähes tuntematon. Toivottavasti teoksen suomennos edistää alan tutkimusta ja opetusta suurin harppauksin. Kemian tiedekasvatuksessa työ on jo alkanut, sillä ensimmäinen keminformatiikan kemian opetuksen opinnäytetyö aloitettiin kevään 2020 aikana. Tässä yhteydessä on tärkeää huomata, että vaikka keminformatiikkaa ei ole käsitetasolla kemian opetuksessa aikaisemmin käsitelty, se on ollut integroituneena alan tutkimukseen esimerkiksi Helsingin yliopiston kemian opettajankoulutusyksikön molekyylimallinnustutkimuksen kautta jo yli 20 vuoden ajan.

Tämä teos luo ymmärrystä keminformatiikan perusteisiin ja mahdollistaa siten sen systemaattisen kouluttamisen ja tutkimisen. Teos esittelee keminformatiikan perusteet suomeksi, mikä tekee aiheeseen perehtymisen vaivattomaksi laajalle yleisölle. Teos soveltuu alan opiskelijoille, tutkijoille, teollisuudessa työskenteleville asiantuntijoille ja ennen kaikkea opettajille. Professori Wild käsittelee aihetta siten, että asiat ovat ymmärrettäviä sekä kemisteille että muille luonnontieteilijöille.

Pidän tärkeänä myös sitä, että keminformatiikan opetuksen lisäksi käännös edistää suomen kielen kehittymistä tieteen kielenä. Käännöstyön aikana olen luonut keminformatiikan keskeisille termeille niiden suomenkieliset vastineet. Termityö ei ollut helppoa ja olen saanut käsitteiden suomentamiseen apua monelta henkilöltä. Kiitos avusta Hermanni Pernaa, Anne Pernaa, Juha Oikkonen, Theo Kurten, Markus Metsälä, Topias Ikävalko ja Maija Aksela.

Käännöstyö kesti noin 3,5 vuotta. Se on pitkä aika noin 4 kuukauden työlle. Aloitin projektin vuonna 2017 työskennellessäni vielä kustannusalalla, ja jatkoin sitä nykyisen työn ohessa Helsingin yliopiston kemian osaston yliopistonlehtorina. Käännös on tehty huolella, mutta pääosa siitä on tehty oman työn ohessa iltapuhteena. Projektin aikana olen ymmärtänyt, että kääntäminen on

todella haastavaa asiantuntijatyötä. Kunnioitukseni ammattikääntäjiä kohtaan on korkea.

Teoksen tuottamisessa on otettu huomioon sekä elektronisen PDF-version että painetun version vaatimukset. Esimerkiksi kustannussyistä lähes kaikki kuvat on uudelleen piirretty mustavalkoisiksi, ja linkit liitetty sivukohtaisesti alaviitteisiin. Käännöstyön aikana kaikki linkit on päivitetty tai korvattu toimivilla. Toivottavasti ne ovat mahdollisimman pitkäikäisiä.

Alkuperäisteos on melkein kymmenen vuotta vanha, joten osa sen sisällöstä on jo vanhentunutta. Olen ottanut vapauden kommentoida näitä kohtia alaviitteiden suomentajan kommentti -kategoriassa. Esimerkiksi kymmenessä vuodessa tietokantojen sisältämät yhdistemäärät ovat kaksinkertaistuneet.

Hyvä lukija. Näihin sanoihin päättäen toivon, että nautit käännöksestä. Se jaetaan ilmaiseksi Helsingin yliopiston Helda Open Books -palvelun kautta lisenssillä CC BY-NC-ND, jotta keminformatiikan osaaminen kehittyisi mahdollisimman nopeasti. Kaikki palaute suomennokseen liittyen on erittäin tervetullutta, ja sen voi lähettää osoitteeseen johannes.pernaa@edumendo.fi.

Helsingissä 2.2.2021
FT Johannes Pernaa

Teoksen kirjoittaja – professori David Wild on erittäin arvostettu keminformatiikan vaikuttaja. Hänellä on yli 25 vuoden kokemus alan opetuksesta ja tutkimuksesta.

David Wild johtaa yhtä harvoista keminformatiikkaan keskittyvistä koulutusohjelmista Indianan yliopistossa. Tutkimuksessa hän on erikoistunut laajamittaiseen tiedonlouhintaan sekä kemiallisen ja biologisen informaation yhdistämiseen.

Tässä oppaassa David Wild johdattaa lukijan keminformatiikkaan esittelemällä alan keskeisimmät perusteet. Hän on sisällyttänyt teokseen keminformatiikan ydinalueita, kuten 2D- ja 3D-rakenteiden visualisoiminen tietokoneella, kemiallisen informaation tallentaminen tietokantaan ja sen esiintyminen verkossa ja tieteellisissä julkaisuissa. Lisäksi hän esittää katsauksen keminformatiikan historiaan ja ohjelmistokehitykseen sekä käy läpi alan tunnetuimpia sovelluksia, kuten esimerkiksi QSAR-menetelmän ja molekyylien telakoinnin.

Tavoitteena on käsitellä tutkimusalueen perusteet ytimekkäästi, jonka pohjalta lukija pystyy syventämään osaamistaan itsenäisesti. Teos soveltuu sekä yliopistossa että teollisuudessa työskenteleville kemisteille, biotieteilijöille ja tietojenkäsittelytieteilijöille, jotka ovat kiinnostuneita keminformatiikasta.

Teoksen on suomentanut FT Johannes Pernaa
Helsingin yliopiston kemian osastolta.

ISBN 978-952-94-4390-1 (nid.) | www.edumendo.fi | ISBN 978-952-94-4391-8 (PDF)